Making Mountains of Morality from Molehills of Virtue: Threat Causes People to Overestimate their Moral Credentials

Daniel A. Effron

London Business School

Citation:

Effron, D.A. (2014). Making mountains of morality from molehills of virtue: Threat causes people to overestimate their moral credentials. *Personality and Social Psychology Bulletin, 40*(8), 972-985. doi: 10.1177/0146167214533131

Author Note

I thank Lisa Shu and Adam Waytz for valuable feedback, and Alexander Ananiadis-Basias and Paula Sjökvist for research assistance.

Address correspondence to Daniel Effron, London Business School, Regent's Park, London, NW14SA, England. Email: deffron@london.edu

Abstract

Seven studies demonstrate that threats to moral identity can increase how definitively people think they have previously proven their morality. When White participants were made to worry that their future behavior could seem racist, they overestimated how much a prior decision of theirs would convince an observer of their non-prejudiced character (Studies 1a–3). Ironically, such overestimation made participants appear more prejudiced to observers (Study 4). Studies 5–6 demonstrated a similar effect of threat in the domain of charitable giving – an effect driven by individuals for whom maintaining a moral identity is particularly important. Threatened participants only enhanced their beliefs that they had proven their morality when there was at least some supporting evidence, but these beliefs were insensitive to whether the evidence was weak or strong (Study 2). Discussion considers the role of motivated reasoning, and implications for ethical decision-making and moral licensing.

Keywords: moral credentials, meta-perceptions, moral licensing, racial prejudice, charitable giving, self, moral standards, attribution, ethics, motivated reasoning, threat

Making Mountains of Morality from Molehills of Virtue:
Threat Causes People to Overestimate their Moral Credentials

On trial for ordering the 1995 Srebrenica Genocide, in which thousands of Muslims were murdered, former Bosnian leader Radovan Karadzic claimed that his actions could not be classified as genocide because he holds no anti-Muslim prejudice. As proof, he pointed to the fact that his former barber was Muslim (Blair, 2012). Many observers were unconvinced; it seems that Karadzic overestimated how much his choice of barber gave him "moral credentials" (Monin & Miller, 2001).

The present research examines how the motivation to defend against threats to one's moral character can bias estimates of how others will judge one's past actions. It is often ambiguous how diagnostic of moral character a particular action is. Does giving a dollar to a homeless person prove that one is generous? Does having a Black acquaintance prove that one is not racist? I propose that people are more likely to think that the answer to such questions is "yes" when they experience a threat to their moral identity – and as a result, they are more likely to overestimate how much they have convinced impartial observers of their morality.

People experience moral identity threats not only when they have been accused of wrongdoing, as with Karadzic, but also when they anticipate acting in a way that could call their virtue into question – such as when they expect to behave unethically, to comply with the morally dubious demands of an authority figure, or to undertake legitimately motivated action that could be misconstrued as unethical. Such anticipated threats, I propose, can lead people to view their past actions as if through a telescope – one that makes molehills of virtue look like mountains of morality.

Reinterpreting Past Behavior as Moral Credentials

The present research was motivated by work showing that people feel more comfortable acting in ethically questionable ways when they can point to evidence that they have a virtuous character – a phenomenon termed *moral licensing* (for reviews, see Merritt, Effron, & Monin, 2010; Miller & Effron, 2010). For example, people become more willing to act in ways that could seem racist when they have previously demonstrated a lack of prejudice (Bradley-Geist, King, Skorinko, Hebl, & McKenna, 2010; Effron, Cameron, & Monin, 2009; Effron, Miller, & Monin, 2012; Mann & Kawakami, 2012; Monin & Miller, 2001). More generally, when people reflect on good deeds they have done in the past, they become more likely to make unethical choices (Conway & Peetz, 2012; Jordan, Mullen, & Murnighan, 2011; Mazar & Zhong, 2010; Sachdeva, Iliev, & Medin, 2009). Acting virtuously seems to make people feel that they have moral credentials (Monin & Miller, 2001) – evidence of a virtuous disposition – which give them license to act less than virtuously. Whereas this prior research examines the consequences of believing that one has moral credentials, the present research examines how threat can lead people to construe a behavior as providing credentials in the first place.

My findings complement research showing that people strategically seek moral credentials when they anticipate needing them in the future (Bradley-Geist, et al., 2010; Merritt et al., 2012). For example, White participants expressed stronger support for a Black job applicant, presumably in order to earn non-racist credentials, when they were made to anticipate that their future behavior could seem racist (Merritt, et al., 2012). Whereas these earlier studies revealed a behavioral strategy for addressing moral identity

threats (i.e., enacting credentialing behaviors), the present research examines a cognitive strategy: reconstruing past behaviors as moral credentials.

Meta-Perceptions and Motivated Reasoning

People seek moral credentials in part to appear moral to themselves (Miller & Effron, 2010; Monin & Miller, 2001), but they also care about appearing moral to others (Pillutla & Murnighan, 1995). To determine whether they are projecting a moral image, people must estimate what their behavior signals about their moral standing to others. Research on moral credentials has neglected such estimates, or *meta-perceptions*. Although studies have shown that in some situations observers are willing to grant actors moral credentials (Effron & Monin, 2010; Polman, Pettit, & Wiesenfeld, 2013), actors' beliefs about observers' willingness to do so have largely gone unexamined, and no work has tested the accuracy of these beliefs.

I predict that threats to a moral identity will lead to inflated meta-perceptions of one's moral credentials. How would this occur? Such threats should motivate people to reassure themselves that their moral standing is secure. To seek such reassurance, people may use motivated reasoning to exploit ambiguity about how much their prior behavior says about their moral character (Sloman, Fernbach, & Hagmayer, 2010). Under threat, actors may evaluate their behavior by asking themselves, in effect, "Could this *possibly* be diagnostic of morality?" Actors not under threat, as well as observers, are likely to evaluate the same behavior more dispassionately by asking themselves, "Is this *probably* diagnostic of morality?" – a comparatively more difficult question to answer affirmatively (cf. Dawson, Gilovich, & Regan, 2002; Gilovich, 1991). In other words, threat should lead people to lower their evidentiary standards for concluding that they have moral

credentials (Ditto & Lopez, 1992; Lord, Ross, & Lepper, 1979). Yet people are likely to be unaware of this motivated process (Balcetis, 2009), to underappreciate the subjectivity of the resulting self-perceptions (L. Ross & Ward, 1996), and to use these self-perceptions as a basis for estimating how others see them (Cameron & Vorauer, 2008; Carlson, Vazire, & Furr, 2011; Kenny & DePaulo, 1993). The result, I predict, is that threat will make actors more likely to overestimate how diagnostic their prior moral behavior will seem to observers.

Present Research

I describe seven studies examining how anticipated moral identity threats can lead people to form inflated meta-perceptions of their moral credentials. Studies 1a-4 focused on the domain of racial prejudice. Many contemporary Americans view being non-racist as an important aspect of morality and are thus motivated to secure a non-racist identity before acting in ways that could seem prejudiced (Crandall & Eshleman, 2003; Effron, et al., 2012; Merritt, et al., 2012). Studies 1a and 1b demonstrate that threat can affect meta-perceptions of one's "non-racist credentials," Study 2 establishes boundary conditions, Study 3 shows that threat can lead to overestimation of non-racist credentials, and Study 4 demonstrates that such overestimation can ironically make one seem prejudiced. Studies 5 and 6 examine the effect of threat on meta-perceptions in a different moral domain (charitable giving), and reveal a theoretically important moderator: individual differences in the importance of moral identity.

Studies 1a and 1b

Inflating Non-Racist Credentials

In prior research, foregoing opportunities to make racist judgments licensed participants to express less racially sensitive views (Effron, et al., 2012). Studies 1a and 1b examined whether the threat of seeming prejudiced in the future could lead to more self-flattering meta-perceptions of a previous non-racist judgment. These studies followed identical procedures, except Study 1b tested the hypothesis more conservatively by incentivizing participants to form accurate meta-perceptions.

Method

Participants. White participants, recruited from a university-maintained, US-wide subject pool, completed the study online in exchange for a chance to win a $50 gift certificate. These and all subsequent studies screened out observations from duplicate IP addresses (suggesting multiple completions from the same participant) or that were traced to non-English-speaking countries (raising concerns about comprehension). Seven participants in each of Studies 1a and 1b did not respond to any of the meta-perception items and thus could not be analyzed. I excluded participants who failed at least one comprehension-check (*n*s = 18 and 39 in Studies 1a and 1b, respectively), or did not make the expected non-racist choice, described below (*n*s = 3 in each study; a necessary exclusion to ensure that all participants formed meta-perceptions of the same behavior). The final sample size was 107 and 106 in Studies 1a and 1b, respectively (across studies, 72% female, M_{age} = 40.86, *SD* = 15.42).[1] Exclusions did not differ significantly by condition in either study, $\chi^2_s(1) < .65$, *p*s > .42. (Without exclusions, tests of the

[1] At .8 power for key tests, Studies 1a-4 could detect modest effects (detectable *d*s = .55, .55, .45 .34, and .43, respectively), and Studies 5 and 6 could detect small effects (detectable f^2s = .05 and .03, respectively).

hypothesis were marginally significant in each study, and significant when meta-analyzing both studies).

Procedure. Both studies had three main phases: (1) Participants were induced to make a non-racist choice, (2) they completed a manipulation of anticipated threat, and (3) they estimated how an observer would interpret their earlier non-racist choice.

Non-racist choice. All participants first had an opportunity to make a non-racist choice over a racist one (Effron, et al., 2012): They read a description of a minor theft, viewed photographs of one White and one Black male suspect, examined evidence about each suspect, and indicated whom they thought the criminal was. The thief's race was unspecified, but his description contained some details stereotypically associated with African-Americans (e.g., listening to rap music). The evidence, however, unequivocally pointed to the White suspect. As noted, I retained only participants who chose the (clearly guilty) White suspect – a non-racist choice, but one that is ambiguously diagnostic of a non-racist disposition.

Threat manipulation. Next, participants were shown two statements and told that later, they would have to choose the one they thought was truer and list reasons why it could be true. In the *anticipated threat* condition, both statements compared Blacks unfavorably to Whites (e.g., "Most blacks are more likely to be criminals than whites"). In the control condition, analogous statements compared teenagers unfavorably to people over age 30. Prior research suggested that only the prospect of affirming the negative statements about Blacks would make participants concerned with seeming racist (Effron et al., 2012, Study 6; see also Bradley-Geist et al., 2010).

Measures. Before choosing a statement (but after being warned that they must do

so later), participants completed the measures. As mentioned, I used three comprehension-check questions to identify and exclude inattentive participants (e.g., recalling whether they would be required to write vs. read about one vs. both of the statements).

For the DV, participants answered three questions about their earlier (non-racist) choice: (1) "Suppose someone wanted to know about how prejudiced or unprejudiced you were towards black people. How informative would they find your choice of suspect?" (*not at all*, *slightly*, *somewhat*, *very*, *extremely,* coded 1-5), (2) How much would a separate group of research participants "learn about your feelings towards black people" if they viewed the suspect choice? (*nothing, a very small amount, a moderate amount, a decent amount, a lot,* coded 1-5), and (3) What would others "think about your racial attitudes based on your choice of suspect?" (*they would think I felt [extremely negatively*, *negatively*, *slightly negatively*, *neutral*, *slightly positively*, *positively*, *extremely positively] about black people,* coded 1-7). The three items were standardized and averaged to form a scale measuring meta-perceptions of moral credentials (αs = .74 and .69 in Studies 1a and 1b, respectively). In Study 1b, I reworded these items so that participants estimated how their suspect choice would be rated by a randomly selected subject pool member, and I incentivized accurate meta-perceptions: Participants would win $5 if their estimates matched this person's ratings.

After selecting which statement (described earlier) to write about, participants in both studies completed a manipulation check: indicating how concerned they were that the upcoming writing task would make them "look bad" (1 = *not at all*; 5 = *very*). Then they wrote about the statement, provided demographics, and were debriefed. (At the end of

Studies 1a-4, participants completed an exploratory measure of racial attitudes, which did not consistently moderate the results).[2]

Results and Discussion

Table 1 presents all statistics. As expected, participants in both studies were more concerned about looking bad when they anticipated having to affirm negative characterizations of Blacks (manipulation check). Confirming the hypothesis with an almost identical effect size in both studies, threatened participants thought their choice of suspect would seem significantly more diagnostic of non-racist attitudes than did control participants.[3] These results suggest that the threat of seeming prejudiced in the future can lead people to enhance their meta-perceptions of their non-racist credentials – even in the presence of a financial incentive to form accurate meta-perceptions.

Study 2:

Boundary Conditions

Study 2 sought to establish boundary conditions implied by theories of motivated reasoning, which state that (a) people are unable to jump to desired conclusions without at least some evidential basis (Kunda, 1990; Pyszcynski & Greenberg, 1987), and (b) as long as the desired conclusion has some evidential basis, people's judgments are relatively insensitive to the quality of evidence (Ditto, Scepansky, Munro, Apanovitch, & Lockhart,

[2] Participants in Studies 1a-6 were asked to guess the hypothesis. Only one participant in Study 5 and one in Study 6 speculated that the manipulation was meant to affect responses; results were identical after excluding these participants.

[3] As a late addition to Study 1a, I added a condition in which participants (n = 46 after exclusions) expected to argue that the negative statements about Blacks were false. Meta-perceptions in this condition (M = -.01, SD = .82) were not significantly different than in the threat condition, $t(97) = 1.22$, $p = .23$ – probably because participants in this condition were more worried about looking bad than expected (M = 1.73, SD = 1.07), significantly more so than in the control condition, $t(94) = 2.01$, $p < .05$.

1998). Specifically, Study 2 assessed the extent to which objective data would constrain threatened participants' meta-perceptions of their moral credentials. Participants performed a non-racist behavior, completed the anticipated threat manipulation from Study 1, and then viewed a new manipulation of whether their prior non-racist behavior represented a "large molehill," a "small molehill," or "not even a molehill" of evidence for a non-racist disposition. I predicted that threatened participants' tendency to form inflated meta-perceptions of their non-racist behavior's diagnosticity would be relatively unconstrained by the strength of the evidence, but would be reduced or even eliminated by a lack of any evidence. In other words, I expected threatened participants to make an equivalently sized mountain out of large and small molehills, but not to be able to make a mountain out of nothing.

Method

Participants. Participants were 227 users of Amazon.com's Mechanical Turk service (MTurk) who received $.51. Because MTurk does not easily allow preselection on demographics, the sample was not restricted to Whites, but I dropped Black participants, for whom affirming negative Black stereotypes likely would have a different psychological meaning (n = 16). After dropping participants who did answer any of the meta-perception items (n = 19), failed at least one comprehension-check question (n = 32), or did not make the non-racist choice (n = 3), the final sample size was 173 (61% female; 84% White; M_{age} = 31.34, SD = 11.55). Attrition did not differ significantly by condition, $\chi^2(5) = 7.40$, $p = .19$. (The direction and significance of the results were identical without exclusions, except where noted).

Procedure. The design was a 2 (anticipated threat vs. control) X 3 (amount of evidence for a non-racist disposition: large molehill vs. small molehill vs. no molehill) factorial. Participants completed the tasks from Study 1: They indicated whether they thought a (clearly innocent) Black suspect or a (clearly guilty) White suspect had committed a crime (as noted, those who accused the Black suspect were excluded), they were told that they would later write a negative essay about Blacks (anticipated threat condition) or teenagers (control condition), and they responded to the comprehension checks.

Next, participants imagined that their choice of suspect had been shown to another MTurk user called "J;" depending on randomly assigned condition, J. had ostensibly been told that 20%, 2%, or 0% of MTurk users in previous research had chosen to accuse the Black suspect. To the extent that unusual actions are more readily attributed to dispositions (Jones & Davis, 1965; Kelley, 1973), this manipulation varies whether participants' own accusation of the White suspect should represent a "large molehill" (20%), a "small molehill" (2%), or "not even a molehill" (0%) of attributional evidence for a non-racist disposition from J's perspective. I selected these values because (a) In a pilot test of 122 MTurk participants who estimated what percentage of others would accuse the Black suspect, the median estimate was 20%, and (b) objectively, the 2% condition provides substantially less attributional evidence than the 20% condition, but barely more evidence than the 0% condition, rendering the test of the hypotheses particularly conservative. As a manipulation check, participants indicated how unusual J. would think it was for someone to accuse the Black suspect, referred to as "Suspect #2" (*not at all*, *slightly*, *somewhat*, *very*, and *extremely*, coded 1-5).

I administered the 3-item measure of meta-perceptions from Studies 1a-1b (α = .83), with slight edits to the wording so that the items asked participants' to estimate J.'s perception of their suspect choice. The remainder of the procedure was identical to Study 1a.

I tested two hypotheses. First, the tendency of threatened (vs. control) participants to form inflated meta-perceptions should emerge more strongly when the suspect choice represents at least a molehill of attributional evidence than when it does not. Second, threatened participants should inflate their meta-perceptions to an equivalent extent regardless of whether the behavior represents a "large molehill" or a "small molehill" of evidence.

Results

Manipulation checks. Confirming the success of the threat manipulation, participants expressed greater concern that they would look bad when they expected to write a negative essay about Blacks (M_{threat} = 3.05, SD = 1.53) versus about teenagers ($M_{control}$ = 1.61, SD = .98), $t(171) = 7.45$, $p < .0001$, $d = 1.14$. Also as expected, the evidence manipulation significantly affected participants' beliefs about how unusual an observer would find the accusation of the innocent Black suspect, $F(2, 170) = 14.44$, $p < .0001$, $\eta^2 = .15$: less unusual when 20% versus 2% or 0% of others had ostensibly accused him (respectively, Ms = 3.26, 4.15, and 4.21; SDs = 1.03, .98, and 1.15), Fs(1, 170) = 19.42 and 23.79, ps < .0001, ds = .84 and .90 – but equally unusual when 2% had accused him as when 0% had done so, $F(1, 170) = .10$, $p = .76$, $d = .06$, illustrating how small a molehill of attributional evidence the 2% condition provided.

Meta-perceptions. Inspection of Figure 1 suggests that, consistent with predictions, threatened participants' tendency to form inflated meta-perceptions relative to control participants was equally apparent in the small-molehill (2%) condition and the large-molehill (20%) condition, but absent in the no-molehill (0%) condition (see Table 2 for descriptive statistics). Planned orthogonal contrasts confirmed the specific hypotheses. The first contrast showed that, as predicted, the magnitude of the threat effect was statistically equivalent in the two "molehill" conditions (large-molehill: threat coded +1, control coded -1; small molehill: threat coded -1, control coded +1), $F(1, 167) = .02$, $p = .89$, $d = .03$. The second contrast confirmed that the threat effect was significantly smaller in the no-molehill condition (threat coded +2, control coded -2) than in the two "molehill" conditions (in each, threat coded -1; control coded + 1), $F(1, 167) = 5.45$, $p = .02$, $d = .36$. (Pairwise comparisons showed that the threat effect was significantly larger in the small-molehill condition than in the no-molehill condition, and marginally larger in the large-molehill condition than in the no-molehill condition, Fs[1, 167] = 4.26 and 3.83, ps = .04 and .05, ds = .38 and .36, respectively; without excluding any participants, both these comparisons were marginal).

Additional analyses showed that, consistent with Studies 1a and 1b, threated participants had (marginally) more flattering meta-perceptions of their credentials than control participants in the small-molehill condition and the large-molehill condition, Fs(1, 167) = 3.82 and 3.31, ps = .05 and .07, ds = .53 and .48, respectively (without exclusions, the second test was significant). Combining these two conditions to increase power revealed a significant threat effect, $F(1, 167) = 7.12$, $p = .008$, $d = .51$, 95% CI for mean

difference = [.11, .73]. As predicted, no such threat effect emerged in the no-molehill condition, $F(1, 167) = .88$, $p = .35$, $d = .24$, 95% CI = [-.62, .22].

Discussion

Replicating the results of Studies 1a-1b, Study 2 found that the threat of doing a potentially prejudiced task enhanced participants' belief that their past behavior would convince an observer of their lack of prejudice. Extending Studies 1a-1b, Study 2 found that this effect was virtually identical in size regardless of whether moderate or weak evidence supported this belief, and that the effect was absent when no evidence supported this belief. The pattern shown in Figure 1 suggests that threat led participants to reduce their evidentiary standards for concluding that they had non-racist credentials: Control participants were unmoved by even a large molehill of evidence, whereas only a small molehill of evidence was enough to affect threatened participants. That is, threatened participants reported more self-flattering meta-perceptions when their behavior *could possibly* seem diagnostic of non-racist attitudes, while control participants did not – perhaps because control participants instead based their judgments on whether their behavior *would probably* seem diagnostic (cf. Dawson, et al., 2002; Gilovich, 1991). These results suggest that the desire to believe that one has moral credentials can lead people to make an equivalently large mountain from both large and small molehills of virtue – but at least a molehill is required to make a mountain.

A potential alternative explanation is that foregoing the racist suspect choice may have seemed more diagnostic of non-prejudice when contrasted against the negative statements about Blacks shown in the threat condition. It is unclear, however, that this purely cognitive mechanism would have predicted the specific moderating effect of

evidence strength, which seems more consistent with the motivated-reasoning processes just described. As explained subsequently, Studies 4 and 5 also favor a motivated process.

Study 3

Overestimating One's Non-Racist Credentials

In Studies 1a-2, threatened participants were more likely than controls to believe that an observer would attribute their prior behavior to a non-racist disposition. How accurate were these beliefs? Study 3 addressed this question by comparing participants' meta-perceptions to observers' actual perceptions.

Method

Actors' data were taken from Study 1a, 1b, and from the small-molehill (2%) and large-molehill (20%) conditions in Study 2. A sample of 152 new participants served as observers (72 from the subject pool used in Studies 1a and 1b, and 80 from the one used in Study 2; subject pool did not affect the results). Four observers did not answer any of the DVs, and, as in the prior studies, I dropped participants who identified as Black (n = 8), failed at least one comprehension check (n = 22), or accused the Black suspect (see below; n = 2). The final sample was 116 observers; demographics were comparable to Studies 1a-2. (Results without exclusions are presented below).

Observers were shown the criminal identification task used in Studies 1a-2 and, without accusing a suspect themselves, learned that a randomly selected participant from a prior study had accused the (guilty) White suspect. Consistent with what actors in Study 2 had been told, observers recruited from the Study 2 subject pool received information about the ostensible proportion of prior participants who had accused the Black suspect (either 2% or 20%); this variation did not affect results and is not discussed further.

For the dependent measure, observers used three items, analogous to the meta-perception items used Studies 1a-2, to indicate their own perceptions of how diagnostic the actor's non-racist suspect choice was of his or her racial attitudes: How informative this choice was about the actor's racial prejudice, how much they had learned about this person's feelings towards Black people, and how positive or negative they thought his or her attitudes about Black people were (α = .68). Next, comprehension-checks asked observers to identify which suspect had been accused, as well as the race of each suspect. (The comprehension check was administered before the DVs for about half of participants). Finally, observers were asked whom they personally thought had committed the crime.

Results and Discussion

As shown in Figure 2, actors tended to mispredict how diagnostic an observer would find their suspect choice, but the direction of this misprediction depended on whether actors had been exposed to a moral identity threat. The measure of diagnosticity (i.e., actors' meta-perceptions and observers' actual perceptions) differed significantly among threatened actors (n = 164), control actors (n = 160), and observers (n = 116), $F(2, 437) = 9.64$, $p < .0001$, $\eta^2 = .04$. Threatened actors overestimated how diagnostic an observer would find their suspect choice, $F(1, 437) = 5.39$, $p = .02$, 95% CI for mean difference = [.03, .40], $d = .28$ ($p = .0002$ without excluding any actors or observers). No such overestimation was found among control actors, who showed a marginally

significant tendency to underestimate, $F(1, 437) = 2.81$, $p = .09$, 95% CI = [-.35, .03], $d = .20$ (without exclusions, $p = .85$).[4]

These results demonstrate that worrying about seeming prejudiced in the future can lead people to overestimate how much their past behavior would convince an observer of their lack of prejudice.

Study 4:

Overestimating Non-Racist Credentials Can Appear Prejudiced

When someone points to a trivial act of virtue as evidence for her morality, people may infer that she is insecure, has inappropriately low moral standards, or is "protesting too much" to compensate for a tarnished moral character. Thus, overestimating one's moral credentials may ironically make one seem less moral. Study 4 examined this possibility.

Method

Participants. I recruited 54 White individuals (16 male, 36 female, 2 unknown) from the subject pool used in Studies 1a-1b. One participant did not respond to any of the DVs, and I excluded those who failed at least one attention check (described subsequently; $n = 9$), leaving a final sample of 45. (The direction and significance of the results were identical without exclusions).

Procedure. Participants saw the criminal decision-making task used in the prior studies, and learned that a previous participant ("L.") had chosen "Suspect #1" (i.e., the guilty White suspect). Next, participants read that "L." had answered three questions about this choice (the same three used to assess meta-perceptions of prejudice in Studies

[4] The marginal underestimation is consistent with the actor-observer difference, for which support is mixed (Jones & Nisbett, 1972; Malle, 2006).

1a-2). Participants did not answer these questions themselves, but instead indicated whether L. would "seem more prejudiced" if his answers overestimated versus underestimated how informative of racial attitudes people found his suspect choice, how much they thought they had learned from his choice about his racial attitudes, and how positive they thought his feelings towards black people were based on his choice (overestimation coded 1; underestimation coded -1; forced-choice; order of responses randomized; three items averaged, $\alpha = .72$). Participants then indicated which suspect L. chose and tried to recall that suspect's race (comprehension checks).

Results and Discussion

Positive values on the DV indicate a belief that overestimating non-racist credentials seems more prejudiced, and negative numbers indicate the reverse. The scale mean was significantly greater than 0 ($M = .30$, 95% CI = [.07, .53], $SD = .11$), $t(44) = 2.59$, $p = .01$, $d = .39$, indicating that participants on average thought that overestimation would seem more prejudiced than underestimation. Indeed, two-thirds of participants thought that overestimation would seem more prejudiced (i.e., 30 out of 45 had a score > 0).

Studies 1a-3 showed that participants responded to a threat to a non-prejudiced self-image by overestimating their non-racist credentials. Study 4 suggests that, ironically, such overestimation may make individuals seem more prejudiced than if they had estimated their credentials more conservatively.

Study 5

Inflating Moral Credentials from Charitable Giving

Study 5 assessed generalizability by examining whether a threat to one's moral self-concept could affect meta-perceptions of one's prior charitable behavior. Study 5 also examined the role of threat more directly by testing mediation by a measure of anxiety – a feeling associated with the experience of threat (Spencer, Steele, & Quinn, 1999).[5] Finally, Study 5 tested a theoretically important moderator: individual differences in the centrality of moral identity to the self-concept (Aquino & Reed, 2002). People high in this trait (i.e., *high moral-identifiers*) respond more defensively to moral identity threats (Mulder & Aquino, 2013). If inflating meta-perceptions of moral credentials is a self-defense strategy, then a moral identity threat should have the biggest effect on the meta-perceptions of high moral-identifiers.

Participants learned that they either would or would not be taking a test that often revealed unconscious moral character flaws (threat manipulation). I hypothesized that the prospect of taking the test would spark anxiety (mediator) – particularly among high moral-identifiers (moderator). This anxiety, in turn, should lead to more flattering meta-perceptions of a trivially charitable choice that participants had made earlier (DV). (See Figure 3, top panel).

Method

Participants. Participants were 206 MTurk users. Seven participants did not answer any of the meta-perception items, and I excluded people who had previously completed a pilot version of the study (n = 2), failed at least one comprehension check (n = 29), or did not make the expected charitable choice (described below; n = 16). The final

[5] It would have been inappropriate in Studies 1a-2 to test the manipulation check (i.e., worry about looking bad) as a mediator because it came after the DV. Theoretically, worry should increase meta-perceptions of credentials, but increased meta-perceptions should subsequently decrease worry; it is unclear what to predict for the net effect.

sample contained 152 participants (70 in the control condition, 82 in the threat condition; 64% male; M_{age} = 32.30, SD = 11.12). Marginally more participants were dropped from the control condition, $\chi^2(1) = 3.38$, $p = .09$, but the direction and significance of results were identical when excluded participants were retained, except where noted.

Moral identity. Moral identity centrality, the hypothesized moderator, was measured with the 5-item *internalization* subscale ($\alpha = .81$) of the Aquino and Reed (2002) moral identity measure. Participants indicated how important it was for them to have nine moral characteristics (e.g., honest, generous, fair; sample item: "I strongly desire to have these characteristics"). Participants also completed the 5-item *symbolization* subscale, a measure of self-presentational aspects of moral identity ($\alpha = .85$). As expected, symbolization did not significantly moderate the results, and is not discussed further.

Charitable choice. Next, all participants chose which of two tasks they would ostensibly complete later in the study: a "visual attention test," which involved memorizing a 13-digit number and searching a 1,600-character matrix for specific letters, or "the charity game" (inspired by http://freerice.com), in which participants would raise $.05 for charity for each of 10 easy trivia questions they answered correctly. (At the end of the study, participants learned that they would not complete either task). As noted, I only retained participants who chose the charitable task (i.e., almost everyone). This choice is ambiguously diagnostic of morality: It raises only a small amount of money and was intended to sound easier and more enjoyable than the alternative.

Manipulation. All participants then read a (bogus) *Science* article describing a highly valid test of "Implicit Moral Character" (IMC) – a trait said to predict ethical

behavior. The article emphasized that because IMC is unconscious, test-takers are frequently shocked to learn how unethical they are. In the anticipated threat condition, participants were told that they would later take the test and learn their score; in the control condition, they were told instead that they would examine another person's score without taking the test themselves. (In reality, participants did neither).

Measures. After answering three filler questions (e.g., whether they had completed the IMC test before) and three multiple-choice comprehension checks about the article (e.g., what the IMC test purports to measure), participants indicated how they felt about the upcoming task (i.e., either taking the test or examining someone's else's score on it, depending on condition): nervous, apprehensive, and worried (response options, coded 1-5: *not at all*, *slightly*, *somewhat*, *very*, and *extremely*). These three items, which were interspersed with fillers (i.e., interested, engaged, bored, and excited), were averaged into a measure of *anxiety,* the hypothesized mediator ($\alpha = .84$).

Finally, participants completed the meta-perceptions measure (DV), for which they imagined that a randomly selected MTurk user learned whether they had previously chosen the charity game or the visual attention task. Then they estimated how informative this person would find this choice if he or she wanted to know (1) how virtuous and (2) how ethical a person they are (*not at all*, *slightly*, *somewhat*, *very*, and *extremely*, coded 1-5), and how much this person would think he or she had learned about (3) their moral character and (4) their generosity (*nothing*, *a very small amount*, *a moderate amount*, *a decent amount*, *a lot*, coded 1-5). I averaged these four items into a meta-perceptions scale ($\alpha = .94$). After providing demographics, participants were debriefed.

Results

Skewness and outliers. To reduce positive skewness in the anxiety measure (skewness = 1.35, $p < .0001$) and to reduce the influence of an outlier, defined as > 3.29 *SD*s away from the mean (i.e., $p < .001$, the cutoff recommended by Tabachnick & Fidell, 2007), I applied a natural log transformation. To reduce negative skewness in the moral identity centrality measure (skewness = -1.69, $p < .0001$) and to reduce the influence of 3 outliers, I squared responses to this measure. No observations were outliers after the transformations.

Moderated path analysis. I predicted that the threat manipulation would arouse more anxiety in high moral-identifiers than in low moral-identifiers, and that such anxiety would lead to increased meta-perceptions of moral credentials (Figure 3, top panel). To test these predictions, I standardized moral identity centrality (the moderator), mean-centered anxiety (the mediator), created an effect code for condition (+1 = threat; -1 = control), and followed the moderated mediation procedure recommended by Preacher, Rucker, and Hayes (2007, Model 2).

Results supported the hypothesized causal path. (Figure 3, bottom panel, shows standardized path coefficients). The threat manipulation increased anxiety among high-moral-identifiers more than among low-moral-identifiers, as shown by a significantly positive threat x identity-centrality interaction, $b = .07$, $t(148) = 2.13$, $p = .04$, $\eta^2_p = .03$. Anxiety, in turn, was positively associated with meta-perceptions, $b = .47$, $t(147) = 2.26$, $p = .03$, $\eta^2_p = .03$. Most importantly, the indirect (mediated) effect of threat on meta-perceptions via anxiety was significantly stronger for high identifiers than for low identifiers, $b = .06$ [.004, .19], as indicated by a bias-corrected bootstrapped 95% CI that did not include 0 (shown in brackets; computed using 5,000 resamples; Edwards &

Lambert, 2007). (This test was marginally significant when no participants were excluded, 90% CI = [.002, .11]). In fact, the indirect effect was significant for high moral-identifiers (1 *SD* above the mean), $b = .09$ [.02, .20], and not for low identifiers (1 *SD* below the mean), $b = .03$ [-.01, .11][6]

In short, the threat manipulation provoked anxiety, which lead to increased meta-perceptions of moral credentials, but only for high moral-identifiers. A caveat is that the total effect of the manipulation on meta-perceptions (i.e., without specifying the causal pathway) was not significant: $b = .06$, $t(148) = .78$, $p = .44$, $\eta^2_p = .004$ for the main effect of the manipulation, and $b = .08$, $t(148) = 1.04$, $p = .30$, $\eta^2_p = .007$ for its interaction with moral identity centrality. However, a meaningful indirect effect does not require a significant total effect of the IV on the DV (Preacher & Hayes, 2008; Rucker, Preacher, Tormala, & Petty, 2011; Shrout & Bolger, 2002; Zhao, Lynch, & Chen, 2010) – in part because tests of indirect effects have greater power than tests of total effects (Rucker, et al., 2011).

Discussion

The data showed the predicted causal pathway: The prospect of taking a morality test sparked anxiety, which led participants to believe that an observer would grant them greater moral credentials for having chosen a fun, charitable task over a boring, non-charitable task. As predicted, this pathway was stronger among individuals whose moral identity is central to their self-concept. These results build on Studies 1-4 by (a) showing that experimentally manipulated feelings associated with the experience of threat predict increased meta-perceptions, and (b) identifying a moderator associated with the

[6] The indirect effect was also significant at the mean of the identity centrality scale, $b = .06$ [.003, .13].

motivation to protect against moral identity threats. A limitation, however, is that the total effect of the manipulation on meta-perceptions (a comparatively lower-powered test that does not specify a causal path; Rucker, et al., 2011) was not significant.

Study 6 used a more powerful design to address this limitation. I modified the procedure to make the threat condition more threatening, adapted the measures and the manipulation to target a more specific dimension of moral identity (i.e., compassion), and recruited a larger sample.

Study 6

Inflating Compassionate Credentials

Method

Participants. Participants were 335 MTurk users. I could not analyze participants who did not responds to any of the meta-perception items (n = 41), and I excluded those who previously completed a pilot version of the study (n = 1), who did not choose the "charity game" (n = 29), or who failed at least one attention check (n = 33). The final sample contained 231 people (53% female; M_{age} = 37.10, SD = 13.34). Exclusions did not differ significantly by condition, $\chi^2(1) = 1.24$, $p = .27$, and left 124 people in the threat conditions and 107 in the control condition. (The direction and significance of all tests was identical without exclusions).

Procedure. I adapted the procedure from Study 5. After choosing between the "visual attention test" and "the charity game," participants completed a modified version of the moral identity scale used in Study 5. I shorted the list of moral characteristics to include only those most closely related to compassion (i.e., caring, compassionate, generous, helpful, and kind). Once again, the internalization subscale was the

hypothesized moderator (α = .70); the symbolization subscale, as expected, did not moderate the results and is not discussed further.

In the threat condition, participants wrote about a time in their adult life when they "hurt someone by doing something selfish, uncaring, or mean" and how their victim probably felt. Then they read an article about a (bogus) test of "Implicit Compassion" and learned that they would momentarily take this test and learn their scores. The article explained that test-takers are often surprised to learn how uncompassionate they really are. I expected that the writing task introduced in this study would make the prospect of taking the test even more threatening than in Study 5.

In the control condition, participants instead wrote about a time in their adult life when they "helped someone by doing something caring, compassionate, generous, or kind" and how their beneficiary probably felt. Then they read the same article as participants in the threat condition, but learned that they would momentarily examine another participant's test score – a less threatening prospect than taking the test themselves. The new writing task should make participants feel particularly secure in their compassionate identity (Jordan, et al., 2011).

After filler items and comprehension checks (see Study 5), participants indicated how informative an observer would find their choice of the "charity game" if this observer wanted to know how caring, compassionate, generous, helpful, and kind they were, and how much the observer would think he or she had learned about their moral character. I averaged these six items to form the meta-perceptions measure (α = .96).[7] This study

[7] Participants also indicated how bad, disappointed, and unhappy with themselves they would feel if the test showed that they were less compassionate than they thought (5-point scales). On average, they expected to feel moderately negative (M = 2.63, SD = 1.19).

omitted Study 5's anxiety measure to address any concerns that it influenced responses to the dependent measures. Finally, participants learned that they would not actually be taking or examining scores on the compassion test.

Results and Discussion

I hypothesized that the threat manipulation would increase meta-perceptions of compassion among high moral-identifiers. As in Study 5, I first reduced negative skewness (-1.45, $p < .0001$) and reduced the influence of 4 outliers on the moral identity measure by squaring the scores. Based on the criterion used in Study 5 (Tabachnick & Fidell, 2007), however, the four extreme scores still represented outliers on the transformed scale. I thus performed a robust regression analysis using an MM-estimator – a technique less prone to biased estimation than standard (OLS) regression and other robust alternatives when multiple outliers on an IV are present (Verardi & Croux, 2009; Yohai, 1987). For this analysis, I regressed meta-perceptions on the manipulation (1 = threat, -1 = control), the transformed moral identity measure (standardized), and their interaction.

The hypothesized interaction (plotted in Figure 4) was significant, $b = .20$, $t(227) = 2.15$, $p = .03$ (neither main effect was significant, $ps > .18$).[8] Consistent with predictions, tests of simple slopes showed that the threat manipulation led high moral-identifiers (1 *SD* above the mean of the transformed moral identity centrality scale) to form more self-flattering meta-perceptions of their choice of the "charity game," $b = .30$, $t(227) = 2.32$, p

[8] With standard (OLS) regression, the level of significance for the interaction is $p = .08$ including the outliers, and $p = .05$ excluding the outliers. In both cases, the significance levels of the simple slopes are the same as reported in the main text.

= .02, β = .30, and that the manipulation had no such effect on low identifiers (1 *SD* below the mean), b = -.11, $t(227)$ = .87, p = .38, β = -.11.

In sum, the more powerful design of Study 6 seems to have enabled detection of an effect that was only observed indirectly through a mediating variable in Study 5: A threat to participants' compassionate identity led to enhanced meta-perceptions of their prior charitable choice, but only if a compassionate identity was relatively central to their self-concept. This interaction was also significant in a meta-analysis of Studies 5 and 6, $t(379)$ = 2.29, p = .02 using the robust regression technique described earlier. Together, Studies 5 and 6 extend my findings to a moral domain other than racial prejudice. Moreover, the moderation by moral identity centrality – an individual difference associated with the motivation to protect against moral identity threats (Mulder & Aquino, 2013) – supports the idea that motivation plays a key role in this process.

A potential alternative explanation for Study 6's results is that the charitable choice seemed more diagnostic when contrasted against the hurtful behaviors participants wrote about in the threat condition versus the helpful behaviors they wrote about in the control condition. (This alternative cannot account for the Study 5 results, because Study 5 omitted the writing task). This perceptual contrast mechanism, however, has difficulty explaining why moral identity centrality moderated the results – unless individuals high (vs. low) in identity centrality described hurtful behaviors that were more negative and helpful behaviors that were more positive. To test this possibility, two independent coders, blind to participants' identity centrality, used two items to rate the behaviors participants described: extremely hurtful to extremely helpful, and extremely mean to extremely kind (-3 to +3). I averaged ratings across items (αs > .81) and coders (α = .92).

(Eleven responses were uncodable). Participants' identity centrality (transformed measure) did not significantly predict coders' ratings of the helpful behaviors, $r(123) = .10$, $p = .25$, or hurtful behaviors, $r(104) = .04$, $p = .68$. (Analogous analyses using an MM-estimator produced the same results). Thus, the data were not consistent with perceptual contrast.

General Discussion

The present research demonstrates that threats to different aspects of moral identity can lead people to enhance their meta-perceptions of their past moral behavior. An anticipated threat to a non-racist identity led participants to overestimate how much their prior behavior would convince an observer of their lack of racial prejudice (Studies 1a-3). Ironically, such overestimation made them seem more prejudiced to observers than a more conservative estimate would have made them seem (Study 4). In a different domain, a threat to a compassionate identity increased participants' estimates that their previous charitable choice would signal a compassionate disposition – but only if a compassionate identity was particularly important to them (Study 6). Together, these results suggest that the need for moral credentials can make people more likely to think that they have already established such credentials in the eyes of others. The behaviors about which participants formed meta-perceptions arguably represented molehills of virtue at best (i.e., declining to accuse a clearly innocent Black criminal suspect; choosing a fun task that raised $.50 for charity instead of a boring task that raised nothing). Yet threat led participants to treat these molehills more like mountains of morality.

I have argued that this effect is driven by the motivation to defend the self from moral identity threats. Perceptual contrast provides a potential alternative explanation for

Studies 1a-2: The non-racist behavior may have seemed more diagnostic when contrasted against the negative statements about Blacks shown in the threat condition. Several findings, however, favor a motivational mechanism. First, consistent with theories of motivated reasoning, Study 2 found that threatened participants' tendency to enhance their meta-perceptions of a non-racist behavior was relatively unconstrained by the strength of the supporting evidence (cf. Ditto, et al., 1998), but was eliminated by a lack of any evidence (cf. Kunda, 1990). It is unclear that perceptual contrast would have predicted this interaction pattern. Second, Study 5 participants reported heightened meta-perceptions of a charitable behavior only to the extent that the prospect of taking a morality test sparked anxiety, a feeling associated with the experience of threat (Spencer, et al., 1999). Third, the effects in Studies 5 and 6 were driven entirely by individuals who are particularly motivated to protect their moral identities (Mulder & Aquino, 2013). Perceptual contrast has difficulty explaining these effects, which were predicted based on people's motivation to defend against identity threats.

Theoretical Contributions and Future Directions

Whereas prior research showed that people sometimes grant moral credentials to themselves (e.g., Monin & Miller, 2001) as well as others (Effron & Monin, 2010; Polman, et al., 2013), the present studies are the first to test actors' ability to assess their own credentials from observers' perspective – and the first to reveal how threat compromises the accuracy of these assessments.

The present research sheds new light on the motivated use of moral standards. People selectively apply moral standards to support desired conclusions (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009), they set lower moral standards for themselves than

for others (Kruger & Gilovich, 2004; Valdesolo & DeSteno, 2007), they exploit ambiguity in moral standards to rationalize their misconduct (Mazar, Amir, & Ariely, 2008; Schweitzer & Hsee, 2002; Shalvi, Dana, Handgraaf, & De Dreu, 2011; Von Hippel, Lakin, & Shakarchi, 2005), and they strategically forget moral standards that would emphasize their ethical failings (Shu & Gino, 2012). The present research suggests that when people anticipate needing evidence of their morality, they expect their prior behavior to be judged against lower moral standards and thus to earn them better moral credentials.

This phenomenon may help explain why pointing to even paltry virtues in one's past can license less ethical behavior (e.g., Gneezy, Imas, Brown, Nelson, & Norton, 2012; Monin & Miller, 2001). In prior research, participants acted morally licensed after making costless, hypothetical decisions (Effron, et al., 2012; Khan & Dhar, 2006; Mazar & Zhong, 2010). One interpretation is that people have chronically low standards for what constitutes evidence of their own morality. The present research suggests a different interpretation: Paltry virtues may seem like better evidence of one's morality when one requires a moral license. Additional research is needed to test whether this mechanism contributes to moral licensing effects.

People have multiple strategies at their disposal to cope with moral identity threats (Shu & Effron, in press). To acquire moral credentials, people can act virtuously (Bradley-Geist, et al., 2010; Merritt, et al., 2010; Sherman & Gorkin, 1980; Zhong, Liljenquist, & Cain, 2009), distort their memories of their moral track record (Effron, et al., 2012; Effron, Monin, & Miller, 2013; M. Ross, McFarland, & Fletcher, 1981; Tenbrunsel, Diekmann, Wade-Benzoni, & Bazerman, 2010), or lower their standards for what they imagine will count as credentials (the present research). Future research should

examine how these strategies operate together. On the one hand, lowering standards may reduce people's need to employ other strategies. Why enact new credentialing behaviors, for example, when lowered standards can make past behaviors seem credentialing? On the other hand, people may use multiple strategies simultaneously to maximize the odds of successful ego-defense. Lowering standards may even facilitate the use of other strategies: More opportunities to enact or invent credentialing behaviors exist when even a molehill of virtue counts as a credential.

Concluding Thoughts

These studies demonstrate a novel source of flexibility in people's striving to defend against moral identity threats. Participants inflated their moral credentials when there was a possibility that their future behavior would reflect negatively on their morality. Theoretically, this phenomenon could occur whether people anticipate performing genuinely unethical behavior or legitimately motivated behavior that could seem unethical. When people need evidence of their morality, even molehills of virtue can seem like mountains of proof.

References

Aquino, K., & Reed, A., II. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*(6), 1423-1440.

Balcetis, E. (2009). Where the motivation resides and self-deception hides: How motivated cognition accomplishes self-deception. *Social and Personality Psychology Compass, 2*(1), 361-381.

Blair, D. (2012). Radovan Karadzic hails his own 'tolerant' nature and peacemaking, *The Telgraph*. Retrieved from http://www.telegraph.co.uk/news/worldnews/europe/bosnia/9612520/Radovan-Karadzic-hails-his-own-tolerant-nature-and-peacemaking.html

Bradley-Geist, J. C., King, E. B., Skorinko, J., Hebl, M. R., & McKenna, C. (2010). Moral credentialing by association: The importance of choice and relationship closeness. *Personality and Social Psychology Bulletin, 36*, 1564-1575.

Cameron, J. J., & Vorauer, J. D. (2008). Feeling transparent: On metaperceptions and miscommunications. *Social and Personality Psychology Compass, 2*(2), 1093-1108.

Carlson, E. N., Vazire, S., & Furr, R. M. (2011). Meta-insight: Do people really know how others see them? *Journal of Personality and Social Psychology, 101*(4), 831-846.

Conway, P., & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behavior. *Personality and Social Psychology Bulletin, 38*(7), 907-919.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. *Psychological Bulletin, 129*(3), 414-446.

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated reasoning and performance on the Wason selection task. *Personality and Social Psychology Bulletin, 28*(10), 1379-1387.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and non-preferred conclusions. *Journal of Personality and Social Psychobgy, 63*(4), 568-584.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M., & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychobgy, 75*(1), 53-69.

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytic framework using moderated path analysis. *Psychological Methods, 12*, 1-22.

Effron, D. A., Cameron, J. S., & Monin, B. (2009). Endorsing Obama licenses favoring Whites. *Journal of Experimental Social Psychology, 45*, 590-593.

Effron, D. A., Miller, D. T., & Monin, B. (2012). Inventing racist roads not taken: The licensing effect of immoral counterfactual behaviors. *Journal of Personality and Social Psychology, 103*, 916-932.

Effron, D. A., & Monin, B. (2010). Letting people off the hook: When do good deeds excuse transgressions? *Personality and Social Psychology Bulletin, 36*, 1618-1634.

Effron, D. A., Monin, B., & Miller, D. T. (2013). The unhealthy road not taken: Licensing indulgence by exaggerating counterfactual sins. *Journal of Experimental Social Psychology, 49*, 573-578.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life*. New York: Free Press.

Gneezy, A., Imas, A., Brown, A., Nelson, L. D., & Norton, M. I. (2012). Paying to be nice: Consistency and costly prosocial behavior. *Management Science, 58*(1), 179-187.

Jones, E. E., & Davis, K. E. (1965). A theory of correspondent inferences: From acts to dispositions. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 2, pp. 219-266). New York: Academic Press.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the cause of behavior. In E. E. Jones, E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior*. Morristown, NJ: General Learning Press.

Jordan, J., Mullen, E., & Murnighan, J. K. (2011). Striving for the moral self: The effects of recalling past moral actions on future moral behavior. *Personality and Social Psychology Bulletin, 37*, 701-713.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28*, 107-128.

Kenny, D. A., & DePaulo, B. M. (1993). Do people know how others view them? An empirical and theoretical account. *Psychological Bulletin, 114*(1), 145-161.

Khan, U., & Dhar, R. (2006). Licensing effect in consumer choice. *Journal of Marketing Research, 43*(2), 259-266.

Kruger, J., & Gilovich, T. (2004). Actions, intentions, and self-assessment: The road to self-enhancement is paved with good intentions. *Personality and Social Psychology Bulletin, 30*(3), 328-339.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480-498.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098-2109.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin, 132*(6), 895-919.

Mann, N. H., & Kawakami, K. (2012). The long, steep path to equality: Progressing on egalitarian goals. *Journal of Experimental Psychology: General, 141*(1), 187-197.

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633-644.

Mazar, N., & Zhong, C.-B. (2010). Do green products make us better people? *Psychological Science, 21*, 494-498.

Merritt, A. C., Effron, D. A., Fein, S., Savitsky, K., Tuller, D. M., & Monin, B. (2012). The strategic pursuit of moral credentials. *Journal of Experimental Social Psychology, 48*, 774-777.

Merritt, A. C., Effron, D. A., & Monin, B. (2010). Moral self-licensing: When being good frees us to be bad. *Social and Personality Psychology Compass, 4*, 344-357.

Miller, D. T., & Effron, D. A. (2010). Psychological license: When it is needed and how it functions. In M. P. Zanna & J. M. Olson (Eds.), *Advances in experimental social psychology* (Vol. 43, pp. 117-158). San Diego, CA: Academic Press/Elsevier.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81*(1), 33-43.

Mulder, L. B., & Aquino, K. (2013). The role of moral identity in the aftermath of dishonesty. *Organizational Behavior and Human Decision Processes, 121*(2), 219-230.

Pillutla, M. M., & Murnighan, J. K. (1995). Being fair or appearing fair: Strategic behavior in ultimatum bargaining. *Academy of Management Journal, 38*(5), 1408-1426.

Polman, E., Pettit, N. C., & Wiesenfeld, B. M. (2013). Effects of wrongdoer status on moral licensing. *Journal of Experimental Social Psychology, 49*(4), 614-623.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*, 879-891.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*(1), 185-227.

Pyszcynski, T., & Greenberg, J. (1987). Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis-testing model *Advances in Experimental Social Psychology* (Vol. 20).

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. Reed & T. E. (Eds.), *Values and Knowledge* (pp. 103-135). Hillsdale, NJ: Erlbaum.

Ross, M., McFarland, C., & Fletcher, G. J. (1981). The effect of attitude on the recall of personal histories. *Journal of Personality and Social Psychology, 40*(4), 627-634.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass, 5*(6), 359-371.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners: The paradox of moral self-regulation. *Psychological Science, 20*, 523-528.

Schweitzer, M. E., & Hsee, C. K. (2002). Stretching the truth: Elastic justification and motivated communication of uncertain information. *The Journal of Risk and Uncertainty, 25*(2), 185-201.

Shalvi, S., Dana, J., Handgraaf, M. J. J., & De Dreu, C. K. W. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes, 115*(2), 181-190.

Sherman, S. J., & Gorkin, L. (1980). Attitude bolstering when behavior is inconsistent with central attitudes. *Journal of Experimental Social Psychology, 16*(4), 388-403.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*(4), 422-445.

Shu, L. L., & Effron, D. A. (in press). Ethical decision-making: Insights from contemporary behavioral research on the role of the self. In R. Scott & S. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences*: Wiley.

Shu, L. L., & Gino, F. (2012). Sweeping dishonesty under the rug: How unethical actions lead to forgetting of moral rules. *Journal of Personality and Social Psychology, 102*(6), 1164-1177.

Sloman, S. A., Fernbach, P. M., & Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition, 115*(2), 268-281.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4-28.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics*. Boston: Pearson Education.

Tenbrunsel, A. E., Diekmann, K. A., Wade-Benzoni, K. A., & Bazerman, M. H. (2010). The ethical mirage: A temporal explanation as to why we are not as ethical as we think we are. *Research in Organizational Behavior, 30*, 153-173.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*(6), 476-491.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science, 18*(8), 689-689.

Verardi, V., & Croux, C. (2009). Robust regression in Stata. *The Stata Journal, 9*(3), 439-453.

Von Hippel, W., Lakin, J. L., & Shakarchi, R. L. (2005). Individual differences in motivated social cognition: The case of self-serving information processing. *Personality and Social Psychology Bulletin, 31*, 1347-1357.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics, 15*, 642-656.

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*(2), 197-206.

Zhong, C.-B., Liljenquist, K., & Cain, D. M. (2009). Moral self-regulation: Licensing and compensation. In D. De Cremer (Ed.), *Psychological Perspectives on Ethical Behavior and Decision Making* (pp. 75-89). Charlotte, NC: Information Age Publishing.

Figures

Figure 1. Meta-perceptions of non-racist credentials ($M \pm SE$), by threat and evidence manipulations, in Study 2. Threatened participants' tendency to form inflated meta-perceptions relative to control participants was equally pronounced regardless of whether there was a "small molehill" or "large molehill" of supportive evidence – but was absent when there was "not even a molehill" of evidence.

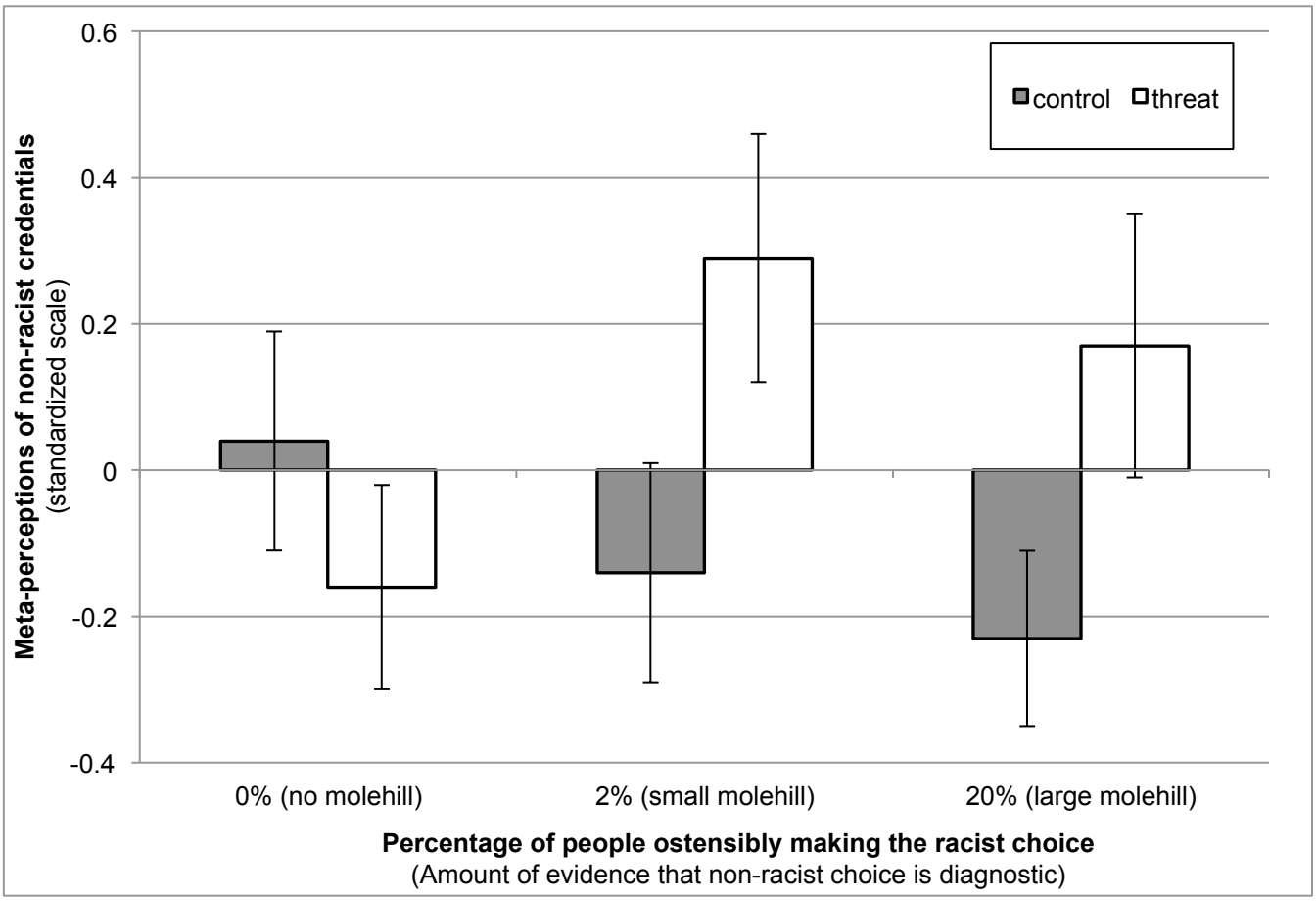


Figure 2. Actors' meta-perceptions of non-racist credentials (Studies 1a-2) compared to observers' actual perceptions of actors' moral credentials ($M \pm SE$). Only threatened actors overestimated their moral credentials.

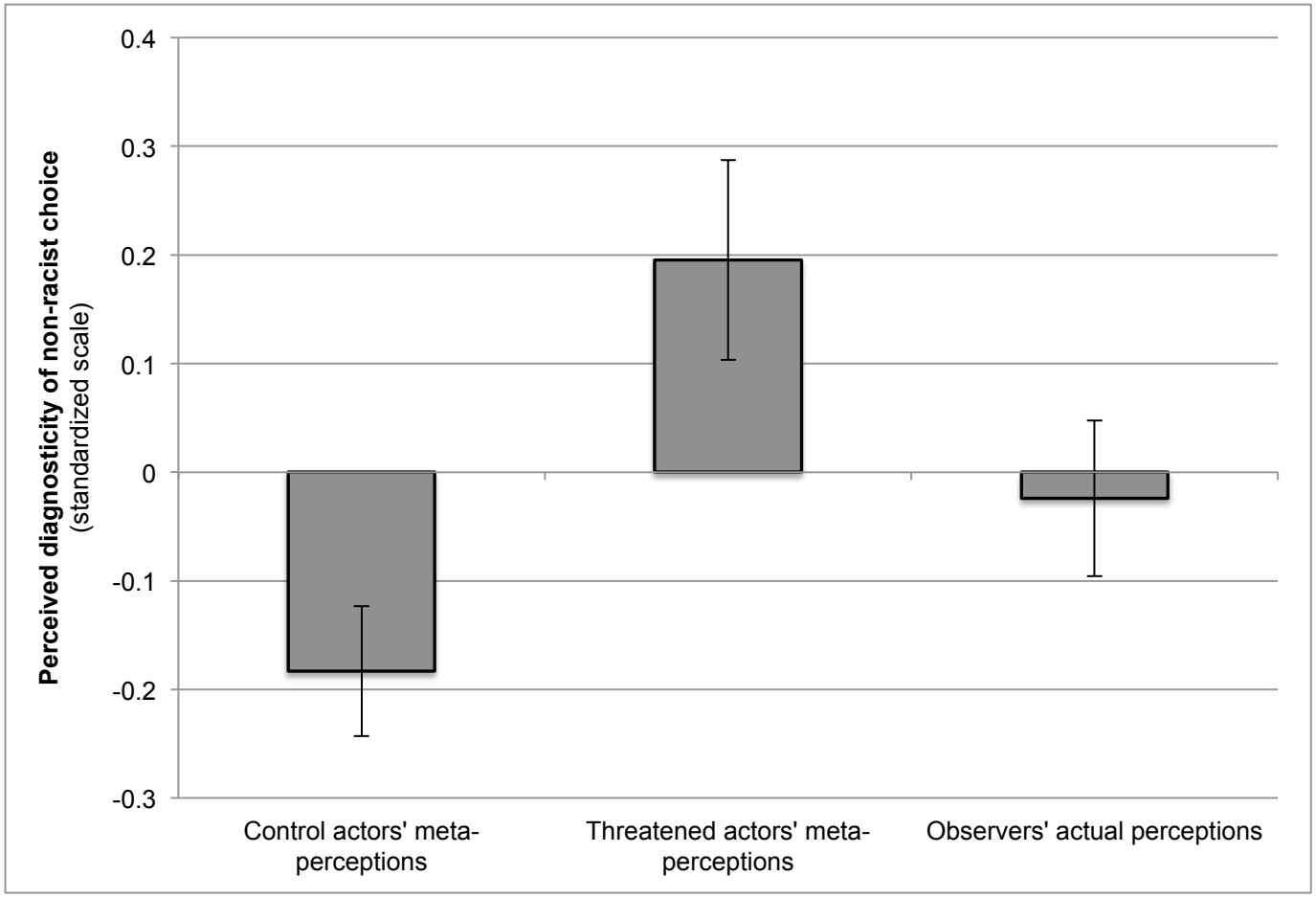
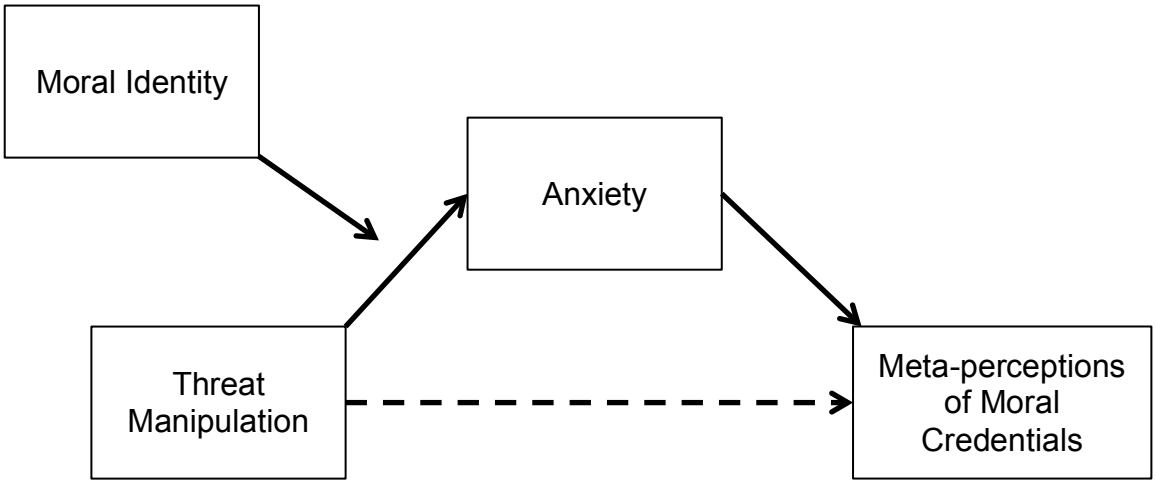
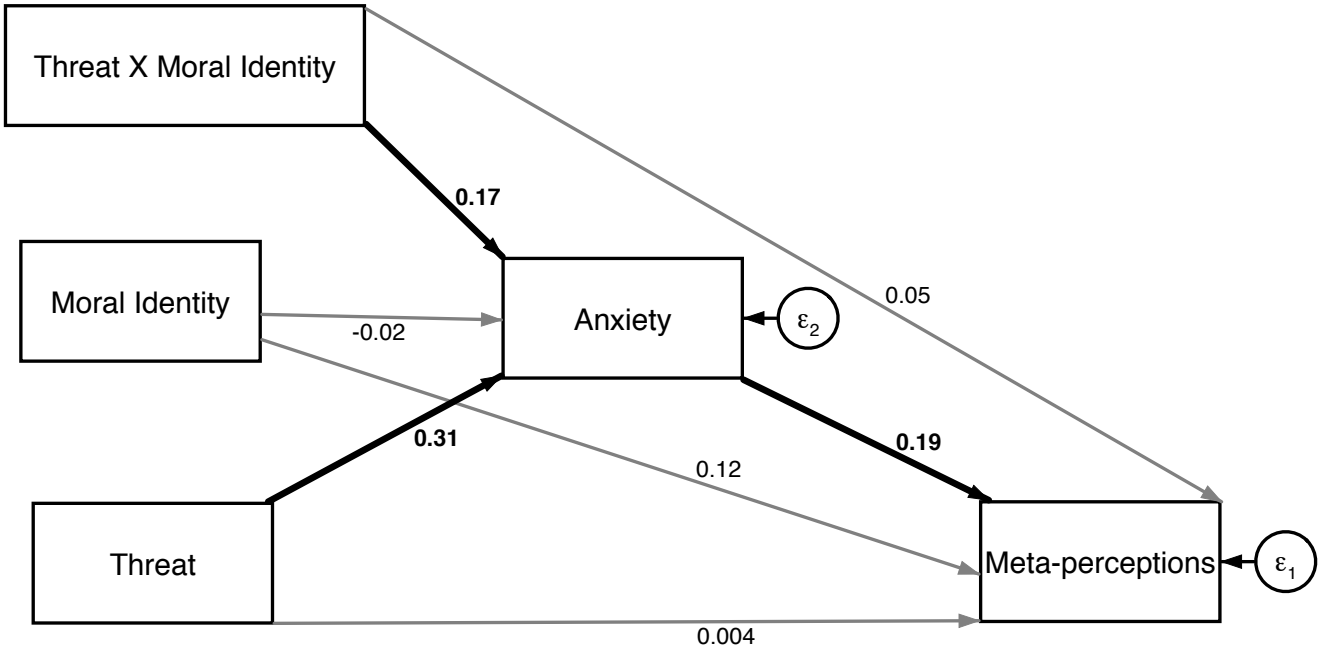


Figure 3. Moderated path analysis for Study 5. As predicted, the indirect effect of threat on meta-perceptions via anxiety was stronger for individuals high in internalized moral identity. *Top panel*: Conceptual model in which moral identity moderates the link between threat and anxiety. *Bottom panel*: Results of path analysis confirming the conceptual model. Values are standardized path coefficients. Bold paths are significant at $p < .05$. Threat condition was coded +1, control condition was coded -1. After applying the transformations to reduce skew described in the main text, moral identity was *z*-scored and anxiety was mean-centered. Covariances among exogenous variables are omitted. Brackets show bias-corrected, bootstrapped 95% CIs for indirect effects.

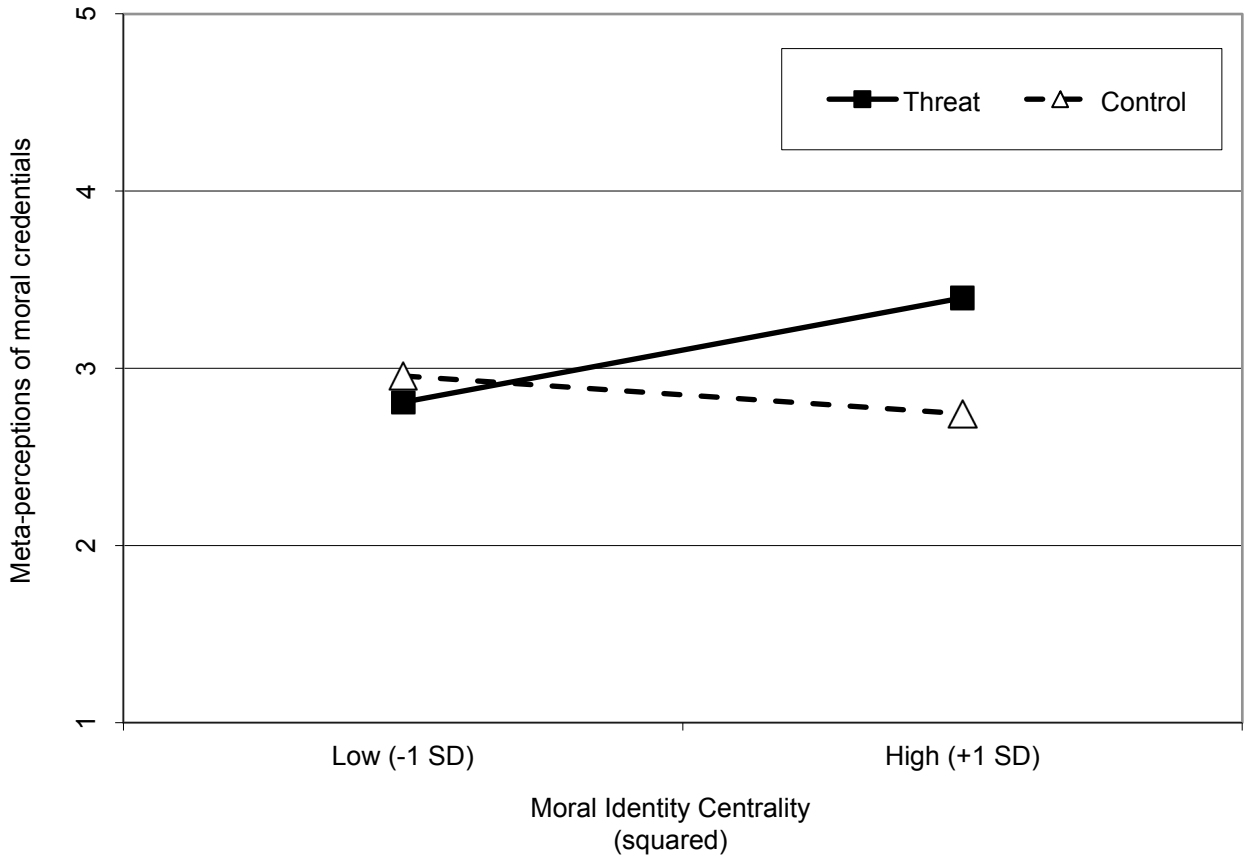




Conditional Indirect Effects (threat → anxiety → meta-perceptions):

- Low moral-identifiers (*M* – *SD*): *b* = .03 [-.01, .11]
- High moral-identifiers (*M* + *SD*): *b* = .09 [.02, .20]
- *Difference (High – low)*: *b* = .06 [.004, .19]

Figure 4. Meta-perceptions of moral credentials in Study 5, by threat manipulation and moral identity centrality. Threat only increased meta-perceptions among high moral-identifiers. The *y*-axis shows the full range of possible responses. Values shown are derived from robust regression analysis.



Tables

Table 1

Results of Studies 1a and 1b

	Condition	*M*	(*SD*)	*n*	95% CI for mean difference	Test of condition difference	*p*	*d*
Study 1a								
Manipulation check	Control	1.37	(0.66)	52	[.78, 1.65]	$t(105) = 5.58$	< 0.0001	1.09
	Threat	2.58	(1.44)	55				
Meta-perceptions	Control	-0.18	(0.69)	52	[.05, .66]	$t(105) = 2.32$	0.02	0.45
	Threat	0.17	(0.88)	55				
Study 1b								
Manipulation check	Control	1.62	(1.02)	53	[.62, 1.49]	$t(104) = 4.79$	< 0.0001	0.94
	Threat	2.68	(1.24)	53				
Meta-perceptions	Control	-0.18	(0.71)	53	[.07, .66]	$t(104) = 2.44$	0.02	0.48
	Threat	0.18	(0.83)	53				

Table 2

Meta-perceptions of non-racist credentials in Study 2: Descriptive statistics

Threat Condition		Evidence Condition: No molehill (0%)	Small molehill (2%)	Large molehill (20%)
Control	*M*	0.04	-0.14	-0.23
	(*SD*)	(.90)	(.76)	(.62)
	n	37	27	28
Threat	*M*	-0.16	0.29	0.17
	(*SD*)	(.72)	(.89)	(.96)
	n	25	27	29