Journal of Experimental Social Psychology 107 (2023) 104461

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Retrospective blind spots in reputation management: Implications for perceived moral standing and trust following a transgression☆

Peter H. Kim [a,*], Alyssa J. Han [b], Alexandra A. Mislin [c], Ece Tuncel [d]

[a] University of Southern California, Marshall School of Business, Dept. of Management and Organization, Hoffman Hall 515, Los Angeles, CA 90089-1421, United States of America
[b] University of Southern California, Marshall School of Business, Dept. of Management and Organization, Hoffman Hall 108, Los Angeles, CA 90089-1421, United States of America
[c] American University, Kogod School of Business, Dept. of Management, 4400 Massachusetts Avenue, NW, Washington, DC 20016, United States of America
[d] Webster University, George H. Walker School of Business & Technology, Department of Management, 470 East Lockwood Avenue, Webster Groves, MO 63119, United States of America

ARTICLE INFO

Keywords:
Ethics
Morality
Reputation
Trust
Impression management

ABSTRACT

Although past research has offered important insights into how people seek to maintain their moral standing, it has generally portrayed this process as a matter of aggregating essentially static interpretations of a target's discrete acts. The present research reveals, however, that such interpretations are often far from static, and that they can change more than targets realize as new events unfold. More specifically, we find that: a) people can discount the diagnostic value of a target's initial deed if that party commits a subsequent act of the opposite valence, b) this occurs when an initial good deed is followed by a bad deed but not when the order is reversed, c) this occurs when evaluating the actions of others but not when evaluating the self, and d) this actor vs. observer difference can ultimately produce divergent beliefs about the target's overall morality, trustworthiness and subsequent trusting behaviors. We also identify a key mediating mechanism for these effects (i.e., the retrospective imputation of nefarious intent). Implications for reputation management, as well as the maintenance and repair of trust, are discussed.

Few reputational concerns play as critical a role in how people navigate the world as the perception of moral character. Moral character has been defined as an individual's disposition to think, feel, and behave in an ethical manner (Cohen & Morse, 2014). It broadly concerns the extent to which a target adheres to standards that others find acceptable (Janoff-Bulman & Carnes, 2013), and it has been found to be more important for impression formation across a wide variety of contexts than traits such as warmth (Goodwin, Piazza, & Rozin, 2014). As such, beliefs about this characteristic have been found to influence an array of social attitudes, behaviors and outcomes. Those who appear to lack moral character, which researchers have also discussed as a perceived lack of integrity (e.g., Kim, Ferrin, Cooper, & Dirks, 2004),[1] have been found to undermine their colleagues' commitment, trust, and inclinations to engage in prosocial behavior (Simons, 2002). Those who lack this characteristic are also punished more harshly for their transgressions (e.g., Laurent, Clark, Walker, & Wiseman, 2014) and find it much more difficult to repair trust after such incidents (e.g., Kim, Cooper, Dirks, & Ferrin, 2013; Kim, Dirks, Cooper, & Ferrin, 2006). Thus, the belief that managers are deficient in this characteristic has been found to harm the stability and effectiveness of the teams they lead

☆ This paper has been recommended for acceptance by Dr. Paul Conway.

* Corresponding author.
E-mail addresses: kimpeter@usc.edu (P.H. Kim), JuRie.Han.2020@marshall.usc.edu (A.J. Han), mislin@american.edu (A.A. Mislin), ecetuncel24@webster.edu (E. Tuncel).

[1] Although moral character and integrity have also been differentiated, the literature currently lacks consensus on how they differ. Some researchers suggest that moral character is broader than integrity because moral character also encompasses traits like consideration of others [which has also been discussed as "benevolence" (Mayer, Davis, & Schoorman, 1995)], honesty-humility, empathic concern, and perspective taking (e.g., Cohen & Morse, 2014). However, other researchers suggest that integrity is broader than moral character because integrity includes additional traits like wholeness, authenticity, word/action consistency, and consistency despite adversity (e.g., Palanski & Yammarino, 2007). Thus, the degree of overlap between these two concepts ultimately depends on how broadly each one is construed.

https://doi.org/10.1016/j.jesp.2023.104461
Received 1 October 2022; Received in revised form 9 February 2023; Accepted 17 February 2023

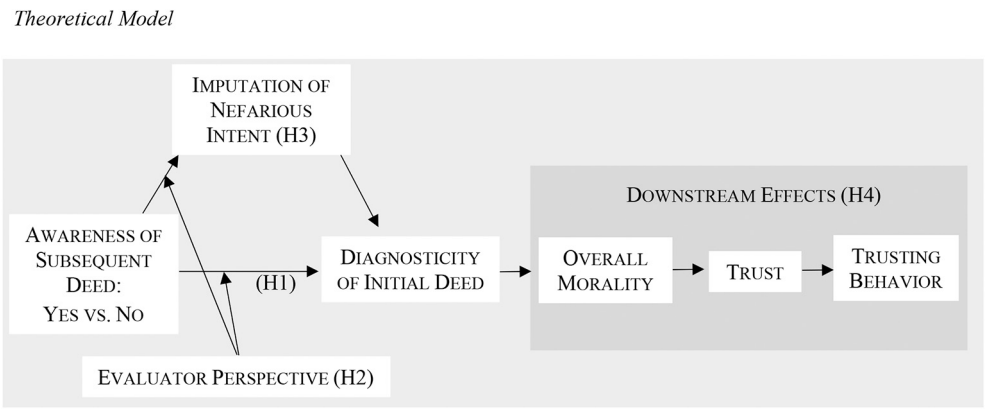

Fig. 1. Theoretical Model.

(Greenbaum, Mawritz, & Piccolo, 2015; Leroy, Palanski, & Simons, 2012; Palanski & Yammarino, 2011). Moreover, organizations that appear to lack integrity similarly suffer, as they draw more ire from investors, activists, and consumers for their transgressions (Janney & Gove, 2010a; Lyon & Maxwell, 2011; Lyon & Montgomery, 2013; Marín, Cuestas, & Román, 2016). For these reasons, a substantial body of research has sought to understand how such beliefs about moral character might ultimately be developed and effectively managed (e.g., Dirks, Kim, Ferrin, & Cooper, 2011; Janney & Gove, 2010b; Kim & Harmon, 2014; Treviño, Hartman, & Brown, 2000).

These issues, furthermore, gain particular significance in light of the fact that few people actually strive to be saints. Rather, individuals have been observed to treat their morality more like a personal bank account to which their ethical and unethical behaviors can add or subtract, respectively (Nisan, 1990). This premise is based on the recognition that although people generally consider morality to be important, they do not require it to be perfect. Hence they tend to act as if they can engage not only in ethical behaviors (which would earn moral credits), but also in unethical behaviors (which would incur moral debits), and ultimately still be considered moral, so long as the balance in that account does not fall below a baseline (e.g., Effron & Monin, 2010; Mazar, Amir, & Ariely, 2008; Mazar & Zhong, 2010). This ethical accounting system has been observed to afford actors the opportunity to engage in behaviors that might otherwise threaten their moral identity and, furthermore, use the exact system when evaluating others to legitimize the allowances they have made for themselves as consistent with generally held principles rather than self-serving ones applicable only to the self (Kim, Wiltermuth, & Newman, 2021).

This model of how to manage one's moral character presumes, however, that the perceived implications of such ethical and unethical behaviors would remain fixed over time to allow these past moral credits and moral debits, respectively, to be aggregated. Yet it is not clear that this assumption is correct. If the perceived implications of those behaviors are not static, this raises the possibility that people's belief that they would be considered moral, so long as the positive implications of their past ethical behavior outweigh the negative implications of their subsequent unethical act, may not hold true. And this could result in serious misjudgments about how they would ultimately be viewed and treated, if one fails to anticipate the extent to which an unethical act would cause one's past ethical behaviors to be discounted by others. The purpose of this paper is to investigate these possibilities by examining: a) how the perceived implications for moral character of one's past ethical and unethical acts might change as new events unfold, b) how such changes might differ for evaluations of the self vs. others, and c) how this might ultimately create divergent beliefs about the target and subsequent trust-relevant behaviors.

1. Theoretical considerations

The basic notion that the implications of people's past behavior for their moral character would remain as a repository of inferences to which new ones are aggregated, lies at the core of research on moral licensing (e.g., Effron & Monin, 2010; Mazar & Zhong, 2010), which has found that individuals' past moral behavior can give them leeway to engage in subsequent transgressions in a different domain (e.g., sexual harassment elicited less condemnation if the target had previously worked to deter adolescent drug abuse).[2] This premise is also consistent with the theory of self-concept maintenance (Mazar et al., 2008), which asserts that people will behave unethically enough to profit but not so unethically that they would be forced to see themselves as unethical. In each case, the findings suggest that people believe the evaluative implications of their past ethical acts would persist as moral credits in their ethical balance sheets to offset the moral debit incurred by their subsequent unethical act, and thus allow them to maintain their moral standing on the whole.

Yet other findings suggest that this basic model of social inference may not always hold true. Research on the reconstructive nature of memory, for example, has found that people's recollections are never a literal account of those experiences. Rather, evidence suggests that these memories are often changed as they are recalled because people tend to supplement other aspects of their knowledge that are unrelated to the actual episode in order to form a more cohesive and well-rounded reconstruction of what happened (e.g., Bell & Loftus, 1989; Loftus & Palmer, 1974). Likewise, research on hypocrisy suggests that when people signal moral values that they subsequently breach, that prior moral stance can be treated, not as a counterbalancing moral credit, but rather as a false signal of moral character that warrants even harsher moral judgments (Jordan, Sommers, Bloom, & Rand, 2017).

These findings are troubling in light of the fact that people's willingness to engage in unethical acts may at least partly arise from the belief that their prior ethical acts serve as a surplus of moral credits that entitles them to engage in subsequent withdrawals (e.g., Effron &

[2] This leeway doesn't entail that past moral behavior will necessarily be followed by a subsequent immoral act. Indeed, whether participants choose to withdraw from their moral bank account (by committing an immoral act) or continue to add to their moral bank account (by committing additional moral acts) has been observed to depend on a host of considerations (e.g., Conway & Peetz, 2012; Mullen & Monin, 2016). The point is simply that past moral behavior *can* liberate people to engage in subsequent immoral deeds in many cases. Indeed, a meta-analytic review of the literature has found that people were at least more likely to follow moral deeds with immoral, unethical, or otherwise problematic behaviors (than with additional moral behaviors) in the majority of studies that review had considered (Blanken, van de Ven, & Zeelenberg, 2015).

Monin, 2010; Mazar et al., 2008; Mazar & Zhong, 2010). Indeed, to the extent that an unethical act leads people to revise their interpretations of what occurred in the past, this suggests that those who rely on their prior moral credits to offset the evaluative implications of a subsequent unethical act may find that those moral credits have essentially vanished right when they are to be used. And if so, those who seek to maintain their moral standing by making sure that any unethical act they commit would be outweighed by their prior ethical behavior, may ultimately be surprised to discover that they are being considered morally bankrupt after all.

However, evidence to support this retrospective possibility is far from conclusive. For example, although research on the reconstructive nature of memory has considered matters of moral judgment (e.g., Galeotti, Saucet, & Villeval, 2020; Kouchaki & Gino, 2015), its focus has been on how people may be motivated to forget their own unethical behaviors rather than whether people would revise their memory of past *ethical* behaviors or whether this would occur when recalling the past deeds of those other than themselves. And though research on hypocrisy has proposed that breaching a moral value one had previously condemned would lead people to treat the initial condemnation as a "false signal" of moral character, that mechanism was simply inferred by comparing differences in participants' overall moral judgments of the target rather than by assessing that mechanism directly (Jordan et al., 2017). Thus, it is unclear whether their findings stem from people revising their interpretations of the initial condemnation or instead simply judging the subsequent moral breach more harshly. Moreover, because that research was specifically focused on explaining people's reactions to blatant forms of hypocrisy, this at least implicitly suggests that this mechanism would only arise when people behave ethically and unethically regarding the exact same kind of deed, rather than apply more broadly to cases where the nature of the good and bad deeds might differ.

Further, even if we accept the premise that people can revise their interpretations of the past as new events unfold, it is not clear how this can be reconciled with the view (from research on moral accounting, moral licensing, and self-concept maintenance) that the implications of one's past deeds would remain as a repository against which subsequent deeds add or subtract. These possibilities have generally not been compared, let alone integrated into a cohesive explanation of what should happen (Kim, Ployhart and Gibson, 2018). Hence, there remains little insight into when or why these different intertemporal models of moral judgment might hold.

We begin addressing this limitation by investigating how support for these competing views, and their implications for moral standing, may depend on both the sequence of events and target of evaluation. Research on the implicit schemas people use when making inferences about others (Reeder & Brewer, 1979) suggests that people intuitively believe that that those with high morality will refrain from unethical behaviors in any situation, whereas those with low morality may exhibit either ethical or unethical behaviors depending on their incentives and opportunities. For this reason, people typically discount a single ethical act as a signal of morality, based on the notion that moral and immoral individuals can each act morally in certain situations (e.g., when there are sufficient incentives for moral behavior or sufficient disincentives to deter immoral behavior). However, people typically consider a single unethical act to offer a reliable signal of immorality, due to the belief that only those who are immoral would behave in unethical ways (e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Ferrin, Kim, Cooper, & Dirks, 2007; Ito, Larsen, Smith, & Cacioppo, 1998; Kim et al., 2004; Tausch, Kenworthy, & Hewstone, 2007). This reasoning is also consistent with research on prescriptive and proscriptive forms of moral regulation (with prescriptive morality focused on "good deeds" that are considered credit worthy and proscriptive morality focused on transgressions that are considered blameworthy), which finds that the costs of failure in the case of proscriptive morality are greater than the rewards of success in the case of prescriptive morality (Janoff-Bulman, Sheikh, & Hepp, 2009). These notions suggest that people will tend to give unethical behavior (negative morality information) more weight than ethical behavior (positive morality information) when making inferences about moral character (Skowronski & Carlston, 1987, 1989). In other words, this literature suggests that people would consider unethical acts to represent larger moral debits, than ethical acts would represent moral credits, in general.

Yet we might also extend this reasoning to determine when perceived differences in information diagnosticity might prompt reinterpretations of prior entries in one's ethical balance sheet as well. In particular, research on cognitive dissonance suggests that people can be discomforted by inconsistencies in the inferences they might make in response to a target's ethical and unethical behaviors, and thus try to resolve those differences (Festinger, 1957). However, rather than doing so simply by adjusting their interpretations of subsequent events to be more in line with their prior inferences, in a manner suggested by the confirmation bias (Nickerson, 1998), we suggest that the nature of this adjustment will also depend on the relative diagnosticity of those prior and subsequent acts. Specifically, to the extent that unethical behaviors are considered more diagnostic of morality than ethical behaviors, a given unethical act should not only be more likely (than an ethical act) to: a) lead people to reevaluate what the actor might have done in the past, but also b) produce inferences that are less likely to be discounted based on events that might follow. And if so, people should ultimately be more likely to revise their prior inferences to better align with new events that unfold when the prior inference concerns an ethical act that is followed by a subsequent unethical act, than when this order is reversed.

We would, furthermore, expect this tendency to weigh negative information about morality more heavily than positive information about morality to depend on the specific target of evaluation. Classic attribution theories suggest that when gauging the extent to which an actor's behavior can be explained by dispositional or situational factors, people attempt to subtract out the effect of the situation and attribute what remains to the individual (Kelley, 1973). However, because societal laws, rules, and norms are generally designed to promote ethical principles (Haidt & Kesebir, 2010; Janoff-Bulman et al., 2009), this generally makes it easier to identify situational explanations for others' ethical than unethical behavior and thus infer that ethical behaviors provide less information (than unethical behaviors) about moral character. Yet evidence suggests that people are not only more aware of the extent to which situational forces affect their own behavior than the behavior of others (Gilbert & Malone, 1995), but also more aware of their own (versus others') intentions (Kruger & Gilovich, 2004). If so, people should: a) be more aware of situational factors that may affect not only their own ethical behavior, but also their own unethical behavior, b) have better insight into their intentions for engaging in those acts, and c) thus give ethical and unethical acts more equal weight when assessing their own moral character than that of others (Kim et al., 2021).

These considerations ultimately suggest that people would consider unethical behavior (negative morality information) to be more informative of moral character when it is committed by others than by the self. In other words, we would expect observers to discount the diagnosticity of an actor's prior ethical behavior, after a subsequent unethical act by that same party, to a greater degree than the actors themselves. And if so, this could create a major blind spot in how actors view the implications of such acts for their moral standing.

Moreover, given that people tend to have better insight into their own (versus others') intentions for such behavior (Kruger & Gilovich, 2004), this may help explain why this self vs. other difference in people's tendency to discount the diagnosticity of an actor's prior ethical behavior (after a subsequent unethical act by that same party) might arise. In particular, when evaluating the self, people are likely to believe that they have ample insight into why they engaged in the prior ethical act and, thus, resist updating that explanation even after engaging in subsequent unethical behavior. However, people's relative lack of

insight into the intentions of others may lead them to wonder if the other's prior ethical behavior was really just an attempt to pave the way for the unethical behavior that party had planned to commit all along. If so, this "retrospective imputation of nefarious intent" may ultimately mediate the extent to which people discount the diagnosticity of a party's past ethical behavior after that party engages in a subsequent unethical act. And finally, to the extent that the inferences people make about these behaviors are eventually combined to reach an overall evaluation of the actor's moral character, we would ultimately expect this tendency to discount an actor's prior ethical behavior, after a subsequent unethical act, to mediate how the actor's moral character is ultimately viewed and treated.

2. Overview of studies

Fig. 1 provides an overview of our theoretical model and predictions, which we tested with five main studies.[3] Study 1 assessed: a) whether people would discount the diagnosticity of a target's initial good deed if they become aware of a subsequent bad deed performed by that individual (H1), b) whether that tendency is stronger when evaluating others rather than the self (H2), c) whether this difference is mediated by the retrospective imputation of nefarious intent for the prior good deed (H3), and d) whether this would in turn affect assessments of the target's overall morality and trust in that party (H4). Study 2 then both replicated and extended these findings by testing whether the actor vs. observer differences from Study 1 are limited to when actors assess their own behavior or would also generalize to actors' expectations of how others would evaluate the actors' behavior. Study 3 next sought to provide a more direct assessment of the proposed mediator. Finally, Studies 4 and 5 sought to replicate and extend the generalizability of our findings with behavioral experiments. In these studies, we report all measures, manipulations, and exclusions. The sample size for each study was determined before any data analysis.

3. Study 1

Study 1 was designed to test Hypotheses 1–4 by comparing how observers would perceive the diagnosticity of an actor's initial good deed with how actors would perceive the diagnosticity of their own initial good deed.

3.1. Method

3.1.1. Participants

Two hundred and fifty-three participants enrolled as undergraduate students at a private university in the eastern part of the United States participated in the study in exchange for course credit.[4] We excluded from analyses participants who failed to correctly answer one or more of the attention check questions. Excluding participants did not affect support for our predictions. The final sample consisted of 197 participants (105 female, 91 male, 1 other). On average, participants were 19.85 years old (ranging from 18 to 26, $SD = 1.35$). We performed a sensitivity power analysis in G*Power 3.1 testing our main effects and interaction effects with a MANOVA, assuming a two-tailed test and an alpha of 0.05. A sample size of $N = 197$ would provide 80% power to detect an effect of Cohen's $f = 0.205$.

3.1.2. Research design

We randomly assigned participants to one of four experimental conditions based on a 2 (awareness of subsequent deed: yes vs. no) by 2 (perspective of evaluator: actor vs. observer) between-subjects design. Participants read a scenario involving analysts working in a demanding environment at an advertising firm. Participants imagined that one of the analysts faced a near-impossible work task with a tight timeline that would likely impact that person's eligibility for an upcoming promotion. Participants then read about a good deed where the target individual Pat, another analyst from a competing team, generously volunteers to help the first analyst with the report, which in turn allows that first analyst to meet the deadline and become a top candidate for promotion. Finally, in the conditions where participants were made aware of the subsequent deed, participants read about a bad deed where Pat compromises a project to which he has been assigned with the first analyst and another team member, by neglecting to follow proper procedures, and tries to cover up his actions in a way that gets the other innocent team member blamed for Pat's actions and ultimately fired.

We varied the perspective that participants were asked to adopt while reading the scenario so that participants were either in the role of the first analyst, as the observer of Pat's good and bad deeds, or the actual actor (in place of Pat) in the scenario engaging in the good and bad deeds themselves. We also varied whether participants assessed the diagnosticity of the good deed immediately following the initial good deed or following the subsequent bad deed, depending on the condition.

3.1.3. Measures

Participants responded to attention check questions following each good or bad deed. We either adapted or developed multi-item scales, depending on their availability in the literature. The new scales were validated with a separate dataset, which revealed reliabilities of 0.86 or above for each.

3.1.3.1. Perceived diagnosticity. We developed a three-item scale to assess the perceived diagnosticity of the initial deed, with minor adjustments to the items to reflect the perspective manipulation.[5] For the Observer conditions, the three items were as follows: (1) Pat's action (helping you with the report) is reflective of his moral character, (2) Pat's action (helping you with the report) genuinely reveals how moral he is, (3) Pat's action (helping you with the report) is informative of how moral of a person he is. Participants rated these items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), with higher values on the scale reflecting greater perceived diagnosticity of the initial good deed ($\alpha = 0.91$).

3.1.3.2. Nefarious intent. We developed a multi-item scale to assess the degree of nefarious intent imputed onto the initial good deed, with

[3] Our early exploratory efforts to develop this project also considered whether judgments of the goodness or badness of the act itself would be more effective at explaining our findings than the perceived diagnosticity of the act. Those exploratory efforts revealed that although results for people's judgments of the act itself were consistent with the results for perceived diagnosticity, those act judgments were ultimately less effective than perceived diagnosticity at explaining the kinds of influences we sought to investigate (e.g., subsequent differences in perceived overall morality). This finding is also consistent with the results of Tannenbaum, Uhlmann, and Diermeier (2011), who found that evaluations of the morality of an act do not necessarily produce corresponding inferences about the moral character of the target person, as well as person-centered accounts of moral judgment, which contend that evaluations of acts themselves (e.g., their consequences or whether a rule has been broken) are ultimately less important for gauging morality than evaluations of what those acts might tell us about the actor's moral character (Pizarro & Tannenbaum, 2012; Uhlmann, Pizarro, & Diermeier, 2015).

[4] We stated in our preregistration (As Predicted #9800, https://aspredicted.org/N5F_P7J) that we would collect data from 250 participants, but we ended up with 253 participants.

[5] The Actor conditions altered the items for this and other scales to ask how the participant, as the Actor, would assess their own behavior [e.g., "Your action (helping with the report) is reflective of your moral character"].

minor adjustments to reflect the perspective. For the Observer conditions, the five items were as follows: (1) Pat had ulterior motives for his action (helping with the report), (2) Pat's action (helping with the report) was strategically motivated for his own personal gain, (3) Pat's action (helping with the report) was an attempt to earn 'brownie points' with his coworkers, (4) Pat's action (helping with the report) was intended to fool others into thinking he is a good person, and (5) Pat's action (helping with the report) was an attempt to disguise his true nature. Participants rated these items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), with higher values indicating greater retroactive imputation of nefarious intent onto the initial good deed (α = 0.85).

3.1.3.3. Overall morality. We assessed the perceived overall morality of the target using a seven-item measure adapted from Helzer and colleagues (Helzer et al., 2014), with minor adjustments to reflect the perspective. For the Observer conditions, the seven items were: (1) Pat is a moral person, (2) Pat does not usually do the right thing (reverse-coded), (3) Pat is not an ethical person (reverse-coded), (4) Pat tries to act in moral ways, (5) Pat is not a moral person (reverse-coded), (6) Pat is an ethical person, and (7) Pat usually does the right thing. Participants rated these items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) (α = 0.93).

3.1.3.4. Trust. We assessed the perceived trustworthiness of the target with a multi-item scale adapted from Kim and colleagues (Kim et al., 2004), with minor adjustments to reflect the perspective. For the Observer conditions, the four items were: (1) Pat is trustworthy, (2) You would not let Pat have any influence over issues that are important to you (reverse-coded), (3) You would keep an eye on Pat (reverse-coded), and (4) You would give Pat a task or problem that is critical to you, even if you could not monitor his actions. Participants rated these items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) (α = 0.86).

3.1.3.5. Trusting behaviors[6]. We developed a multi-item scale to assess expected trusting behaviors towards the target, with minor adjustments to reflect the perspective. In the Observer conditions, participants were asked to imagine they had been promoted to a managerial position and asked to rate their likelihood of engaging in the following trusting behaviors towards the target if given the opportunity: (1) recommend Pat for an important company position, (2) allocate Pat important company resources, and (3) put Pat in charge of handling relations with one of the company's most important clients. In the Actor conditions, participants were asked to indicate how likely they believed their boss would be to engage in these same actions. Participants rated these items on a 5-point scale (1 = extremely unlikely, 5 = extremely likely) (α = 0.95).

3.1.4. Pretests[7]

Prior to conducting the main study, we also conducted three pretests. The first pretest sought to assess whether the predicted tendency to discount the diagnosticity of an initial good deed after becoming aware of a subsequent bad deed performed by that individual could be found in real-life managers. Thus, we tested 42 full-time managers enrolled in an online MBA program and analyzed the subset of respondents that correctly answered all the attention check questions. The final sample consisted of 31 participants (11 female, 20 male), averaging 34.52 years in age. Participants read a scenario with the same good deed as in Study 1 and with a bad deed in which Pat steals items from a coffee shop and blames an innocent barista for the missing items. The results of a repeated-measures ANOVA revealed a significant effect of observers' awareness of the subsequent bad deed on the perceived diagnosticity of Pat's initial good deed, $F(1,30) = 15.616$, $p < .001$, $\eta_p^2 = 0.342$. As expected, participants evaluated Pat's initial good deed as significantly less diagnostic of morality after Pat's bad deed ($M = 3.94$, $SD = 1.61$) than they did before Pat's bad deed ($M = 5.29$, $SD = 1.04$).

Second, we wanted to confirm our theorized boundary condition for our predictions – that the tendency to engage in this discounting would occur when a good deed was followed by a bad deed, but not when this order was reversed. Thus, we randomly assigned 302 MTurk participants to one of four experimental conditions in a 2 (awareness of subsequent deed: yes vs. no) by 2 (ordering of deeds: good before bad vs. bad before good) between-subjects design, based on actual events that occurred between Steve Jobs and Steve Wozniak, the co-founders of Apple, and asked them to assess the perceived diagnosticity of the actor's initial deed using the scale from the main study, with the items adjusted to reflect the type of deed being evaluated (i.e., good or bad) (α = 0.96).[8] As expected, the results revealed a significant awareness x order interaction, $F(1, 261) = 20.006$, $p < .001$, $\eta_p^2 = 0.071$. Participants who evaluated an initial good deed after reading about a subsequent bad deed viewed the initial good deed as significantly less diagnostic of moral character ($M = 3.96$, $SD = 1.46$) than those only informed of Pat's initial good deed ($M = 5.28$, $SD = 1.21$), $F(1, 261) = 49.119$, $p < .001$, $\eta_p^2 = 0.158$, but those evaluating an initial bad deed after reading about a subsequent good deed did not view the initial bad deed differently from those only informed of Pat's initial bad deed, $F(1, 261) = .490$, $p = .484$, $\eta_p^2 = .002$.

Finally, we conducted a third pretest with 253 MTurk participants to assess whether actors and observers would differ in their evaluations of the good or bad deed in Study 1 when either deed was presented on its own.[9] This was done to evaluate the extent to which any actor vs. observer differences we observe could be attributed to a self-serving motivation to view one's own deeds in a more favorable light. We conducted planned contrasts, which did not reveal significant differences in the extent to which actors and observers considered the bad deed to be bad [$t(100.000) = -0.663$, $p = .509$, $d = -0.194$] or diagnostic of moral character [$t(218) = -1.002$, $p = .318$, $d = -0.198$], despite the fact that actors should be particularly motivated to interpret this kind of deed in self-serving ways. And though the analyses revealed significant differences in the extent to which actors and observers considered the good deed to be good [$t(117.965) = -2.147$, $p = .034$, $d = -0.531$] and diagnostic of moral character [$t(218) = -2.179$, $p = .030$, $d = -0.398$], these results indicated that observers actually interpreted the actor's good deed *more* favorably (goodness of good deed, $M = 4.36$, $SD = 0.59$; diagnosticity of good deed, $M = 6.27$, $SD = 0.77$) than the actors themselves ($M = 4.12$, $SD = 0.64$; $M = 5.88$, $SD = 0.99$). Thus, we did not find evidence that self-serving motivations could explain support for our study predictions.

[6] Although this distinction between trust and trusting behaviors is consistent with past research, which has observed that trusting behaviors can arise for a variety of reasons that do not necessarily reflect the presence of trust (Kim et al., 2009), such as if there are sufficient incentives or safeguards to cooperate despite one's lack of trust, these measures are often highly correlated. Hence, even though we have pre-registered each of these measures and their causal relationship, readers may choose instead to view these measures simply as alternative ways of tapping into the same underlying trust construct.

[7] Details about the pretests are provided in the supplemental materials.

[8] As in the main study, we excluded from analyses participants who failed to correctly answer one or more of the attention check questions. We also excluded the 25 participants who recognized the real-life event on which our scenario was based. Excluding these participants did not affect support for our predictions. The final sample consisted of 265 participants (148 female, 116 male, 1 other), averaging 32.72 years in age (ranging from 18 to 84, $SD = 10.82$).

[9] We excluded from analyses participants who failed to correctly answer one or more of the attention check questions. The final sample consisted of 222 participants (92 female, 130 male), averaging 35.61 years in age (ranging from 19 to 72, $SD = 11.69$).

3.2. Main study results

Correlations for the study variables are provided in Table 1. Means and standard deviations by condition are provided in Table 2. Two-way ANOVAs revealed significant main effects of awareness of the subsequent bad deed on nefarious intent [$F(1, 193) = 14.372$, $p < .001$, $\eta_p^2 = .069$], the perceived diagnosticity of the initial good deed [$F(1, 193) = 39.404$, $p < .001$, $\eta_p^2 = 0.170$], overall morality [$F(1, 193) = 156.291$, $p < .001$, $\eta_p^2 = .447$], trust [$F(1, 193) = 174.188$, $p < .001$, $\eta_p^2 = .474$], and trusting behaviors [$F(1, 193) = 124.607$, $p < .001$, $\eta_p^2 = .392$]. Participants who evaluated the target's initial good deed after reading about the target's subsequent bad deed imputed greater nefarious intent onto the initial good deed, perceived the initial good deed to be less diagnostic of the target's moral character, perceived the target to be lower in overall morality and trustworthiness, and expected less trusting behavior towards the target than those who made these assessments immediately after reading about the good deed.

The results also revealed significant main effects of perspective on nefarious intent [$F(1,193) = 21.437$, $p < .001$, $\eta_p^2 = .100$], the perceived diagnosticity of the initial good deed [$F(1, 193) = 13.470$, $p < .001$, $\eta_p^2 = 0.065$], overall morality [$F(1,193) = 29.418$, $p < .001$, $\eta_p^2 = .132$], trust [$F(1,193) = 68.501$, $p < .001$, $\eta_p^2 = .262$], and trusting behaviors [$F(1,193) = 52.070$, $p < .001$, $\eta_p^2 = .212$]. Compared to the actors themselves, observers imputed greater nefarious intent onto the actor's initial good deed, perceived the initial good deed to be less diagnostic of the actor's moral character, perceived the target to be lower in overall morality and trustworthiness, and expected less trusting behavior towards the actor.

However, these main effects were ultimately qualified by significant awareness x perspective interactions (See Fig. 2). We found significant interaction effects for nefarious intent [$F(1, 193) = 7.185$, $p = .008$, $\eta_p^2 = 0.036$], perceived diagnosticity [$F(1, 193) = 4.766$, $p = .030$, $\eta_p^2 = 0.024$], trust [$F(1, 193) = 4.833$, $p = .029$, $\eta_p^2 = 0.024$], and trusting behaviors [$F(1, 193) = 80.988$, $p < .001$, $\eta_p^2 = 0.296$], but not for overall morality [$F(1, 193) = 0.549$, $p = .460$, $\eta_p^2 = 0.003$]. These results revealed that learning about the actor's subsequent bad deed led to a greater imputation of nefarious intent onto the initial good deed, a greater discounting of the perceived diagnosticity of the initial good deed, and a greater decrease in trust and expected trusting behaviors for observers than for actors.

Table 1
Correlations for Study Variables (Study 1).

Variable	1	2	3	4	5
1. Nefarious Intent	–				
2. Diagnosticity	−0.431***	–			
3. Overall Morality	−0.407***	0.586***	–		
4. Trust	−0.414***	0.563***	0.808***	–	
5. Trusting Behaviors	−0.318***	0.491***	0.593***	0.634***	–

*** $p < .001$.

Table 2
Means and Standard Deviations by Awareness of Subsequent Deed x Perspective (Study 1).

			Awareness of Subsequent Bad Deed		
			No *M (SD)*	Yes *M (SD)*	Total *M (SD)*
Perspective	Observer *M (SD)*	Nefarious Intent	2.98 (0.89)	4.05 (1.31)	3.48 (1.23)
		Diagnosticity	5.53 (0.76)	4.30 (1.06)	4.95 (1.10)
		Overall Morality	5.23 (0.81)	3.56 (0.75)	4.45 (1.14)
		Trust	4.58 (0.83)	2.47 (0.75)	3.59 (1.32)
		Trusting Behaviors	4.02 (0.65)	2.02 (0.75)	3.08 (1.22)
	Actor *M (SD)*	Nefarious Intent	2.66 (1.21)	2.84 (1.10)	2.75 (1.16)
		Diagnosticity	5.74 (0.98)	5.15 (1.16)	5.46 (1.11)
		Overall Morality	5.82 (0.75)	4.34 (1.11)	5.11 (1.20)
		Trust	5.42 (0.83)	3.91 (1.26)	4.69 (1.30)
		Trusting Behaviors	3.85 (0.63)	3.63 (0.74)	3.74 (0.69)
	Total *M (SD)*	Nefarious Intent	2.80 (1.09)	3.36 (1.33)	3.07 (1.24)
		Diagnosticity	5.65 (0.89)	4.79 (1.19)	5.24 (1.13)
		Overall Morality	5.56 (0.83)	4.01 (1.04)	4.82 (1.22)
		Trust	5.05 (0.92)	3.30 (1.29)	4.21 (1.42)
		Trusting Behaviors	3.92 (0.64)	2.94 (1.09)	3.46 (1.01)

3.2.1. Moderated mediation analyses

Bootstrapped moderated mediation analyses with 5000 bootstrap samples (PROCESS Model 8; Hayes, 2018) supported the prediction that the tendency to discount the diagnosticity of the initial good deed would be mediated by the extent to which actors vs. observers impute nefarious intent onto the actor's initial good deed (95% CI [0.05, 0.49]), as well as predictions that this differential tendency to discount the diagnosticity of the initial good deed would mediate each of the following dependent variables when tested on its own: overall morality (95% CI [0.03, 0.50]), trust (95% CI [0.02, 0.48]), and trusting behaviors (95% CI [0.01, 0.28]; see Appendix A).

3.2.2. Moderated serial mediation analyses

We also tested several forms of moderated serial mediation (PROCESS Model 85; Hayes, 2018). These analyses provided support for a significant indirect sequential relationship in which perspective moderated the effect of awareness of the subsequent bad deed on the imputation of nefarious intent, which then affected the perceived diagnosticity of the initial good deed, which in turn affected the perceived overall morality of the target (95% CI [0.02, 0.19]), and a significant indirect sequential relationship in which perspective moderated the effect of awareness on the imputation of nefarious intent, which affected the perceived diagnosticity of the initial good deed, which then affected the perceived overall morality of the target, and then affected trust in that individual (95% CI [0.01, 0.12]). We then examined whether these influences would ultimately affect expected trusting behaviors. Although perspective was found to moderate the direct effect of awareness on trusting behaviors, support was not found for a significant indirect sequential relationship in which perspective moderated the effect of awareness on trusting behaviors through the imputation of nefarious intent, the perceived diagnosticity of the initial good deed, the perceived overall morality of the target, and trust (95% CI [−0.0001, 0.02]).[10] Finally, we tested an array of alternative causal orderings to assess the viability of serial mediation models beyond what our research predicted and found minimal support for those alternatives (see Appendix B).

[10] Excluding participants who failed the attention checks proved to be the more conservative test of our prediction. When including all participants in our analyses, we did find support for this moderated serial mediation relationship (95% CI [0.002, 0.02]).

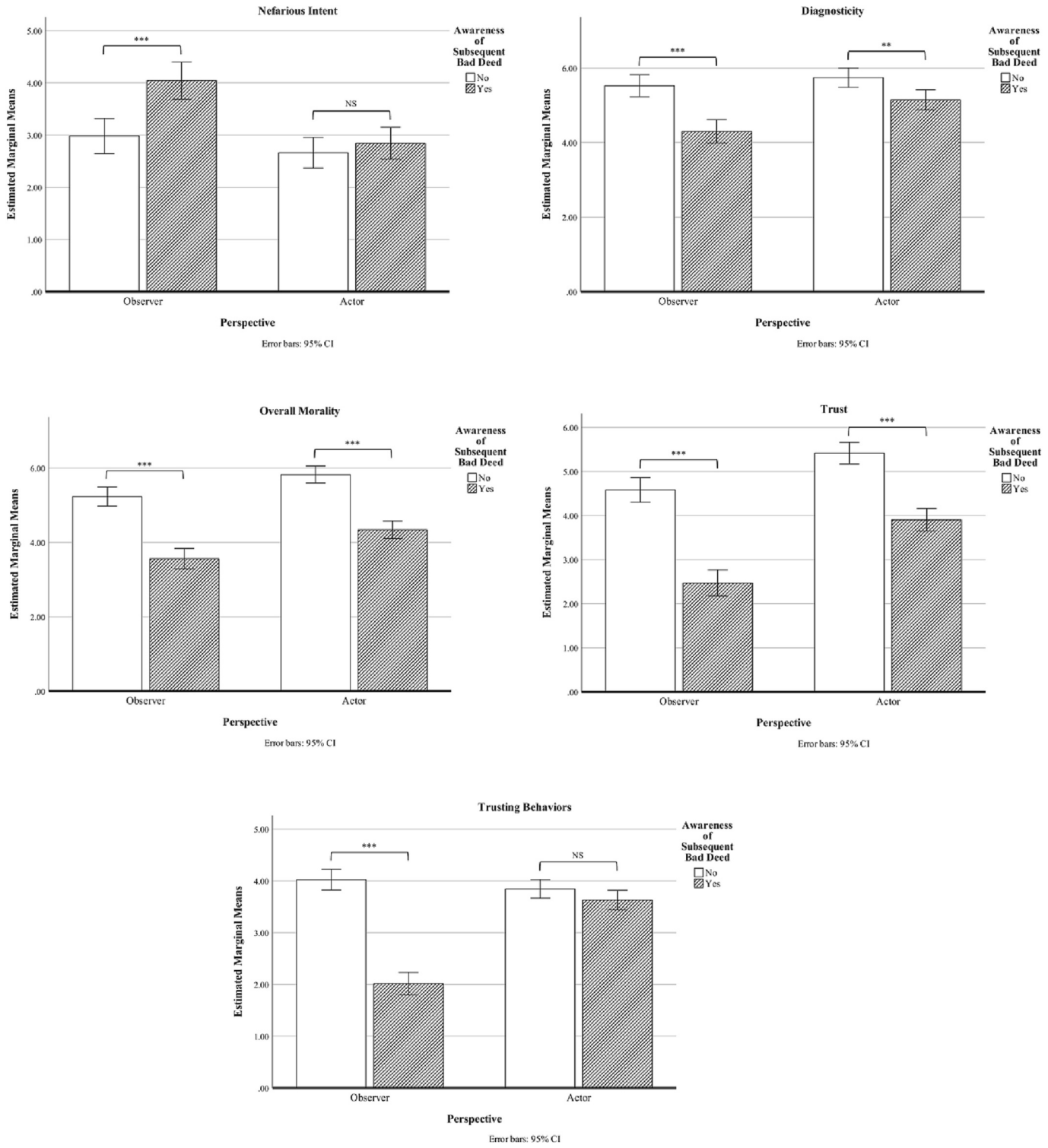

Fig. 2. Means by Awareness of Subsequent Deed x Perspective (Study 1).
** $p < .01$. *** $p < .001$. NS = Not Significant.

3.3. Discussion

The results from Study 1 support our predictions that people would discount the diagnosticity of an individual's initial good deed following a subsequent bad deed performed by that same individual (H1), and that this discounting would occur to a greater degree with observers than the actors themselves (H2). We also found that following the actor's subsequent bad deed, observers imputed greater nefarious intent onto the actor's initial good deed, considered the actor less trustworthy, and expected less trusting behaviors towards the actor than the actors themselves. Moreover, we found support for our prediction that this differential tendency to discount the diagnosticity of the actor's initial good deed would be mediated by the extent to which people retroactively impute nefarious intent onto that initial good act (H3), and that this discounting effect would in turn influence perceptions of the target individual's overall morality, trust in the target individual, and trusting

behaviors (H4).

Study 1 therefore provides strong support for a mismatch between actors and observers in how they evaluate and respond to the same behaviors. Moreover, the combination of these results with those from the third pretest provide little support for a motivational explanation for these findings. Nevertheless, one might argue that actors could still recognize that others would discount the implications of the actors' initial good deeds (after the actor commits a subsequent bad deed), even if actors are less inclined to do so themselves. This seems unlikely, given that if that was the case, Study 1 would not have found actor-observer differences in expected trusting behaviors (wherein actors were explicitly asked to consider how their boss, an observer, would treat them). However, we ran a second study to rule out this possibility more directly and thus provide further support for the notion that actors would misjudge the reputational implications of their own actions, as well as to determine whether stronger support could be found for an awareness x perspective interaction effect on perceived overall morality.

4. Study 2

Study 2 was designed to investigate whether the actor vs. observer differences observed in Study 1 are limited to when actors assess their own behavior or would also generalize to actors' expectations of how others would assess their behavior. To the extent that actors' expectations for how their behavior would be assessed by others do *not* mirror actors' assessments of their own behavior (and are instead more in line with how observers would actually interpret these actions), this could suggest that actors may have sufficient insight into how their behaviors would be perceived to maintain their reputations for moral character. However, given that people tend to believe that they are being fair and consistent in how they account for the ethical implications of their own and others' behavior (Kim et al., 2021), we predict that actors will expect others to make the same inferences about the actors' deeds as actors would themselves. In other words, we expect to find support for the same four predictions from Study 1 (H1, H2, H3, H4) even when comparing observers' assessments with actors' assessments of how others would evaluate the implications of the actors' behavior. If so, this would underscore the potential for actors to misjudge how their prior good deeds would be viewed and thereby fail to maintain their moral standing.

4.1. Method

4.1.1. Participants

Two hundred and fifty-eight participants from the United States were recruited through Amazon's Mechanical Turk (MTurk) and were each paid $0.50 to complete the study.[11] We again excluded from analyses participants who failed to correctly answer any of the attention check questions. Excluding participants did not affect support for our predictions. The final sample consisted of 226 participants (97 female, 129 male). On average, participants were 36.36 years old (ranging from 18 to 70, $SD = 11.63$). We performed a sensitivity power analysis in G*Power 3.1 testing our main effects and interaction effects with a MANOVA, assuming a two-tailed test and an alpha of 0.05. A sample size of $N = 226$ would provide 80% power to detect an effect of Cohen's $f = 0.192$.

4.1.2. Research design

Study 2 used the same 2 (awareness of subsequent deed: yes vs. no) by 2 (perspective of evaluator: actor vs. observer) between-subjects design and the same scenario as Study 1.

[11] We stated in our preregistration (As Predicted #27683, https://aspredicted.org/28R_CMG) that we would collect data from 250 participants, but we ended up with 258 participants.

4.1.3. Measures

Participants responded to the same attention check questions as in Study 1. We also used the same multi-item scales as Study 1 to assess the perceived diagnosticity of the actor's initial good deed ($\alpha = 0.94$), nefarious intent ($\alpha = 0.94$), overall morality ($\alpha = 0.93$), trust ($\alpha = 0.84$), and trusting behaviors ($\alpha = 0.96$), with two modifications. Items for the actor conditions were modified so that rather than having participants assess the implications of their own behavior, they would now indicate how they thought others would assess the implications of that behavior [e.g., "Others would see your action (helping with the report) as reflective of your moral character"]. Additionally, to keep the items more consistent throughout, we modified the trusting behavior items so that participants in the actor conditions were no longer asked to assess their *boss'* likelihood of engaging in trusting behaviors towards them (as in Study 1) but were rather asked to assess *others'* likelihood of engaging in trusting behaviors towards them. Likewise, participants in the observer conditions were no longer asked to imagine they had been promoted to a managerial position but were simply asked about their likelihood of engaging in trusting behaviors towards the actor if they had the opportunity.

4.2. Results

Correlations for the study variables are provided in Table 3. Means and standard deviations by condition are provided in Table 4. Two-way ANOVAs revealed significant main effects of awareness on nefarious intent [$F(1,222) = 4.109$, $p = .044$, $\eta_p^2 = 0.018$], perceived diagnosticity of the initial good deed [$F(1, 222) = 40.831$, $p < .001$, $\eta_p^2 = 0.155$], overall morality [$F(1, 222) = 246.570$, $p < .001$, $\eta_p^2 = 0.526$], trust [$F(1, 222) = 257.324$, $p < .001$, $\eta_p^2 = 0.537$], and trusting behaviors [$F(1, 222) = 264.104$, $p < .001$, $\eta_p^2 = 0.543$]. Participants who evaluated the actor's initial good deed after the actor's subsequent bad deed imputed (or expected others to impute) greater nefarious intent onto the initial good deed, perceived (or expected others to perceive) the initial good deed as less diagnostic of the actor's moral character, evaluated (or expected others to evaluate) the actor as lower in overall morality, reported (or expected others to report) lower trust, and were less willing (or expected others to be less willing) to engage in trusting behaviors towards the actor compared to those who made these assessments immediately after the actor's initial good deed. We did not find a significant main effect of perspective on any of our measures.

However, we did find significant awareness x perspective interaction effects for all of our measures (See Fig. 3), including nefarious intent [$F(1, 222) = 7.537$, $p = .007$, $\eta_p^2 = 0.033$], perceived diagnosticity [$F(1, 222) = 23.258$, $p < .001$, $\eta_p^2 = 0.095$], overall morality [$F(1, 222) = 7.994$, $p = .005$, $\eta_p^2 = 0.035$], trust [$F(1, 222) = 16.699$, $p < .001$, $\eta_p^2 = 0.070$], and trusting behaviors [$F(1, 222) = 15.498$, $p < .001$, $\eta_p^2 = 0.065$]. These findings revealed that the actor's subsequent bad deed led observers to impute greater nefarious intent onto the initial good deed, consider the initial good deed to be less diagnostic, consider the actor to possess less overall morality, and exhibit less trust in and trusting behaviors towards the actor than the actors themselves expected.

Table 3
Correlations for Study Variables (Study 2).

Variable	1	2	3	4	5
1. Nefarious Intent	–				
2. Diagnosticity	−0.271***	–			
3. Overall Morality	−0.457***	0.578***	–		
4. Trust	−0.400***	0.520***	0.883***	–	
5. Trusting Behaviors	−0.131*	0.558***	0.754***	0.790***	–

* $p < .05$. *** $p < .001$.

Table 4
Means and Standard Deviations by Awareness of Subsequent Deed x Perspective (Study 2).

			Awareness of Subsequent Bad Deed		
			No *M (SD)*	Yes *M (SD)*	Total *M (SD)*
Perspective	Observer *M (SD)*	Nefarious Intent	3.31 (1.93)	4.33 (1.44)	3.77 (1.79)
		Diagnosticity	5.96 (0.74)	4.13 (1.53)	5.13 (1.48)
		Overall Morality	5.58 (1.01)	2.92 (1.10)	4.37 (1.69)
		Trust	5.01 (1.18)	2.23 (0.97)	3.75 (1.77)
		Trusting Behaviors	4.20 (0.67)	1.89 (1.03)	3.15 (1.43)
	Actor (Meta-Perceptions) *M (SD)*	Nefarious Intent	4.08 (1.52)	3.93 (1.45)	4.01 (1.48)
		Diagnosticity	5.08 (1.33)	4.83 (1.20)	4.97 (1.28)
		Overall Morality	5.17 (0.98)	3.32 (1.23)	4.34 (1.43)
		Trust	4.62 (0.80)	2.96 (1.17)	3.88 (1.28)
		Trusting Behaviors	3.71 (0.73)	2.30 (0.99)	3.08 (1.10)
	Total *M (SD)*	Nefarious Intent	3.71 (1.76)	4.13 (1.45)	3.90 (1.64)
		Diagnosticity	5.51 (1.17)	4.48 (1.41)	5.05 (1.38)
		Overall Morality	5.37 (1.01)	3.12 (1.18)	4.35 (1.56)
		Trust	4.81 (1.02)	2.60 (1.13)	3.81 (1.53)
		Trusting Behaviors	3.95 (0.75)	2.10 (1.03)	3.11 (1.27)

4.2.1. Moderated mediation analyses

Finally, bootstrapped moderated mediation analysis with 5000 bootstrap samples (PROCESS Model 8; Hayes, 2018) supported the prediction that the tendency to discount the diagnosticity of the initial good deed would be mediated by the imputation of nefarious intent onto the actor's initial good deed (95% CI [0.03, 0.40]), as well as predictions that this differential tendency to discount the diagnosticity of the initial good deed would mediate each of the following dependent variables when tested on its own: overall morality (95% CI [0.34, 0.97]), trust (95% CI [0.25, 0.70]), and trusting behaviors (95% CI [0.25, 0.68]; see Appendix A).

4.2.2. Moderated serial mediation analyses

We also tested several forms of moderated serial mediation (PROCESS Model 85; Hayes, 2018). These analyses provided support for a significant indirect sequential relationship in which perspective moderated the effect of awareness of subsequent bad deed on the imputation of nefarious intent, which then affected the perceived diagnosticity of the initial good deed, which in turn affected the perceived overall morality of the target (95% CI [0.01, 0.13]). We also found a significant indirect sequential relationship in which perspective moderated the effect of awareness on the imputation of nefarious intent, which affected the perceived diagnosticity of the initial good deed, which then affected the perceived overall morality of the target, and then affected trust in that individual (95% CI [0.01, 0.09]). Moreover, we found a significant indirect sequential relationship in which perspective moderated the effect of awareness on the imputation of nefarious intent, the perceived diagnosticity of the initial good deed, the perceived overall morality of the target, trust in the target, and finally trusting behaviors towards that individual (95% CI [0.002, 0.04]). Finally, we tested an array of alternative causal orderings to assess the viability of serial mediation models beyond what our research predicted and only found support for alternative models in which perceived diagnosticity preceded, rather than followed, the perception of nefarious intent (see Appendix B).[12]

4.3. Discussion

The results from Study 2 replicate and extend support for our predictions. Although people discounted the diagnosticity of an individual's initial good deed following a subsequent bad deed performed by that same individual (H1), this discounting occurred to a greater degree with observers than the actors themselves (even when actors estimated how *others* would evaluate the implications of the actors' behavior) (H2). We also found that following the actor's subsequent bad deed, observers imputed greater nefarious intent onto the actor's initial good deed, considered the actor less moral and less trustworthy, and indicated less willingness to engage in trusting behaviors towards the actor than the actors themselves estimated. The support for a direct awareness x perspective interaction effect for overall morality is notable, since that was the one test that fell short of significance in Study 1. Moreover, this difference in actors' and observers' discounting (or expected discounting) of the initial good deed again was not only mediated by the extent to which these parties retroactively imputed (or expected imputation of) nefarious intent onto the actor's initial good deed (H3), but also affected downstream assessments (or expected assessments) of the target individual's overall morality, trust, and subsequent trusting behaviors (H4). These findings provide strong and consistent support for the notion that actors would fail to anticipate this retrospective discounting by others, even when they were explicitly asked to consider the others' perspective. Thus, there is little reason to expect that actors would be able to account for this effect when considering the reputational implications of their own behavior.

5. Study 3

Study 3 was designed to provide a more robust evaluation of the central mechanism for this research by directly manipulating the actor's nefarious intent for the initial good deed and examining how this would affect the dependent variables of interest. We predicted that observers would discount the diagnosticity of the initial good deed to a greater degree when these observers are informed there was nefarious intent underlying the actor's initial good deed than when they are informed nefarious intent was absent. We also predicted that this discounting effect would, in turn, influence perceived overall morality, trust, and expected trusting behaviors.

5.1. Method

5.1.1. Participants

One hundred and eighty-three participants from the United States were recruited through Amazon's Mechanical Turk (MTurk) and were each paid $1.00 to complete the study.[13] As in our prior studies, we excluded from analyses participants who failed to correctly answer one or more of the attention check questions. Excluding participants did not affect support for our predictions.[14] The final sample therefore consisted

[12] See Study 3 for further validation of our theorized mediation sequence.

[13] We stated in our preregistration (As Predicted #91262, https://aspredicted.org/K83_QY3) that we would collect data from 180 participants, but we ended up with 183 participants.

[14] Although we planned to also exclude those who failed the manipulation checks from the analyses, the reverse coding of one of the items led many participants to fail at least one of the two checks, leading to an unusually high number of participants being excluded. Thus, these study results include participants regardless of whether or not they failed the manipulation checks. However, we found virtually identical results even when excluding those who failed the manipulation checks, with the only exception being the trusting behaviors measure, due to the smaller sample size.

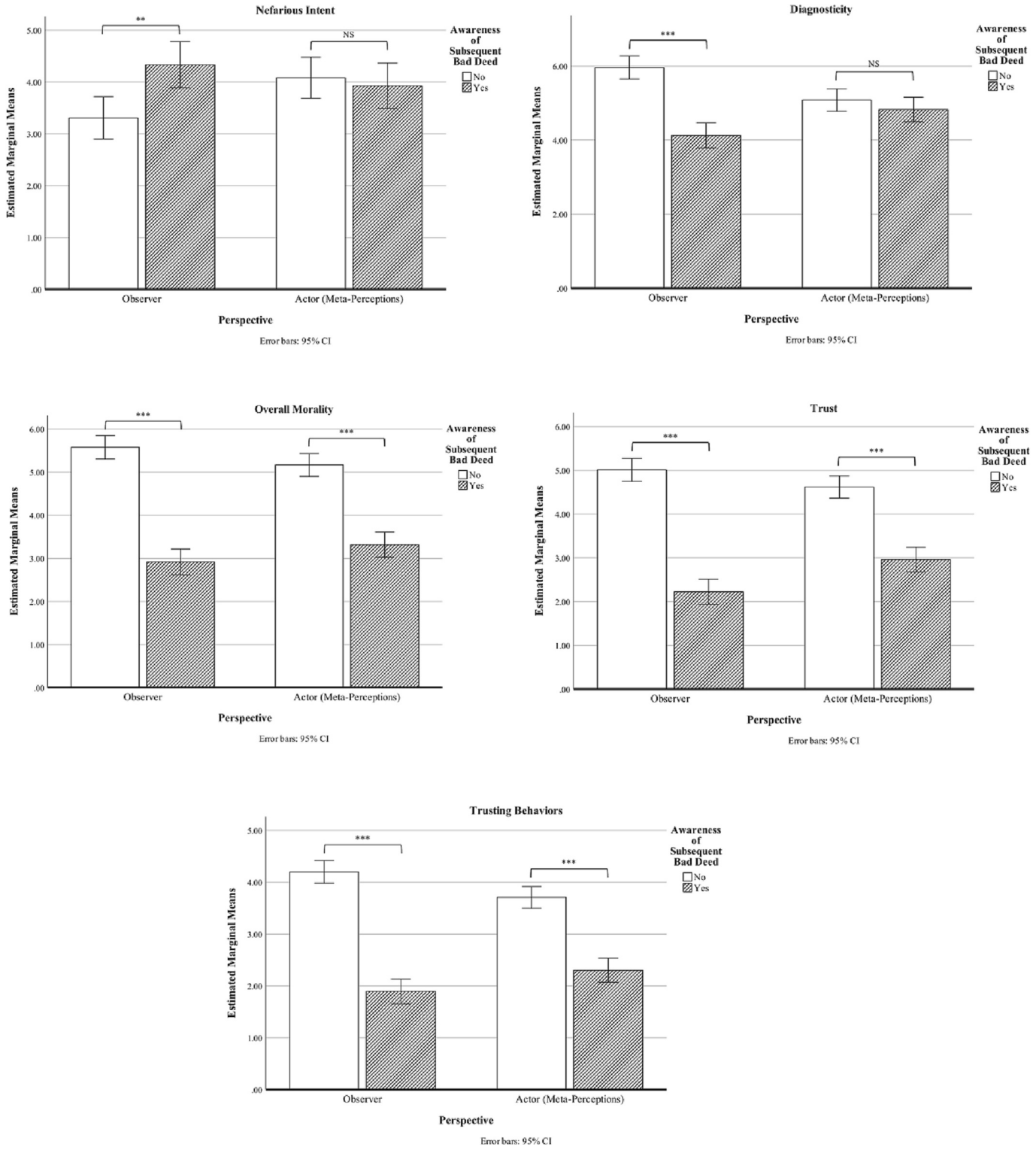

Fig. 3. Means by Awareness of Subsequent Deed x Perspective (Study 2).
** $p < .01$. *** $p < .001$. NS = Not Significant.

of 139 participants (35 female, 104 male). On average, participants were 34.37 years old (ranging from 22 to 72, $SD = 9.32$). We performed a sensitivity power analysis in G*Power 3.1 testing our planned contrasts with a MANOVA, assuming a two-tailed test and an alpha of 0.05. A sample size of $N = 139$ would provide 80% power to detect an effect of Cohen's $f = 0.298$.

5.1.2. Research design

We randomly assigned participants to one of three between-subjects conditions. In our two main experimental conditions, participants read about a subsequent bad deed following an initial good deed that was done either: (1) *without* nefarious intent or (2) *with* nefarious intent. We also included a third control condition in which participants were not made aware of the subsequent bad deed at all. However, because this

condition is not germane to the central purpose of this study, we will focus on the results from our two experimental conditions in the main text and report comparisons with the control condition in a footnote.

Participants were presented with a scenario similar to that of Study 1, but with a few changes. Participants imagined themselves as a consulting analyst at an accounting firm who observes Pat engaging in good and bad deeds. As in Study 1, in the *good deed*, Pat voluntarily helps the participant with a report. In the *bad deed*, Pat is assigned to work on a project with the participant and another analyst, who is the team lead on the project; it later turns out that there was a problem with the financial reporting part of the project that Pat was responsible for completing, and this ultimately leads to the team lead getting fired. The participant later expresses to another coworker, who happens to be close with Pat, that this situation has led them to wonder why Pat previously went out of his way to help them with the report. Specifically, the participant questions whether Pat had an ulterior motive for his initial good deed, such as wanting to make himself look good in front of others. This coworker responds either that they are certain Pat did *not* have ulterior motives since these things are usually kept secret at the firm and no one would find out about what he did (*without* nefarious intent*)* or that they are certain Pat did have ulterior motives since these things are never kept secret at the firm and everyone would eventually find out about what he did (*with* nefarious intent*)*. We therefore manipulated the attribution of nefarious intent onto the initial good deed by informing participants via a third party that Pat either did or did not have ulterior motives for his initial good deed.

5.1.3. Measures

Participants responded to attention check questions following each good or bad deed section of the scenario. Additionally, we included manipulation check questions asking participants whether Pat had ulterior motives for his initial good deed. We then assessed the perceived diagnosticity of the initial good deed (α = 0.75), perceived overall morality (α = 0.71), trust (α = 0.55), and trusting behaviors (α = 0.81) using the same multi-item scales as in the prior studies.

We also included an exploratory measure to ensure that our manipulation of nefarious intent did not create significant differences in people's perceptions of the subsequent bad deed. Participants were asked to assess the extent to which they perceived Pat's bad deed to be bad, and the extent to which they believed Pat was to blame for the outcome of his bad deed (i.e., getting the team lead fired). These two items were rated on a 5-point scale (1 = *not at all*, 5 = *a great deal*), and responses were averaged to compose a scale (r_{SB} = 0.63). Results from this analysis showed that people who perceived Pat as having nefarious intent for his initial good deed did not significantly differ from those who perceived Pat as not having nefarious intent in their perceptions of the subsequent bad deed, $t(84) = 1.029$, $p = .306$, $d = 0.223$. Thus, we did not find evidence to suggest that support for our study predictions could be attributed to differing perceptions of the bad deed itself.

5.2. Results

Correlations for the study variables are provided in Table 5. Means and standard deviations by condition are provided in Table 6. To test our predictions, we conducted one-way ANOVAs, specifically focusing on planned contrasts comparing Condition 1 (*without* nefarious intent) against Condition 2 (*with* nefarious intent). We found significant main effects of nefarious intent on the perceived diagnosticity of the initial good deed [$t(77.942) = -3.217$, $p = .002$, $d = -0.635$], overall morality [$t(136) = -4.005$, $p < .001$, $d = -0.870$], trust [$t(82.622) = -3.620$, $p < .001$, $d = -0.761$], and trusting behaviors [$t(83.966) = -2.094$, $p = .039$, $d = -0.486$].[15] Participants who perceived Pat as having nefarious intent for his initial good deed considered the initial good deed less diagnostic of the actor's moral character, evaluated the actor as lower in overall morality, and exhibited lower trust and trusting behaviors towards the actor than those who perceived Pat as not having nefarious intent (See Fig. 4).[16]

Table 5
Correlations for Study Variables (Study 3).

Variable	1	2	3	4
1. Diagnosticity	–			
2. Overall Morality	0.506***	–		
3. Trust	0.474***	0.721***	–	
4. Trusting Behaviors	0.580***	0.594***	0.641***	–

*** $p < .001$.

Table 6
Means and Standard Deviations by Nefarious Intent (Study 3).

	Nefarious Intent of Initial Good Deed		
	Without Nefarious Intent *M (SD)*	With Nefarious Intent *M (SD)*	Total *M (SD)*
Diagnosticity	5.36 (0.84)	4.56 (1.43)	4.91 (1.27)
Overall Morality	4.51 (0.80)	3.71 (1.02)	4.06 (1.01)
Trust	4.01 (0.80)	3.24 (1.15)	3.58 (1.07)
Trusting Behaviors	3.66 (0.91)	3.19 (1.17)	3.40 (1.09)

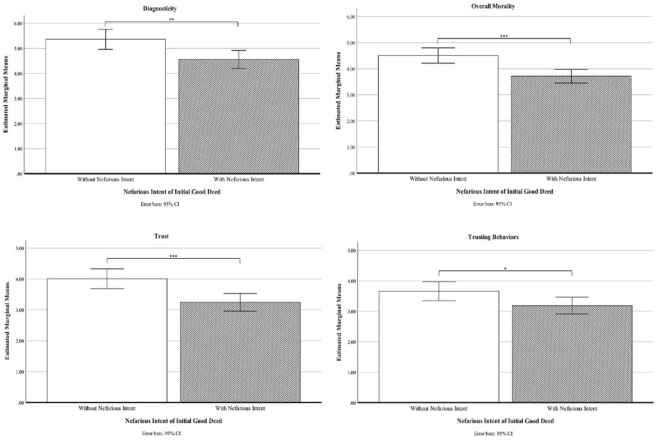

Fig. 4. Means by Nefarious Intent of Initial Good Deed (Study 3).
* $p < .05$. ** $p < .01$. *** $p < .001$.

[15] We found virtually identical results when those who failed the manipulation checks were excluded from the analyses. The only exception was that, due to the smaller sample size, there was no longer a significant effect on trusting behaviors, $t(11.238) = -1.078$, $p = .304$, $d = -0.601$.

[16] We had preregistered (As Predicted #91262) that we would also conduct planned contrasts to compare our control condition (not aware of subsequent bad deed) against our two experimental conditions combined (aware of subsequent bad deed), in order to test for a main effect of awareness of the subsequent bad deed. However, we ultimately realized that the nature of our nefarious intent manipulation in our experimental conditions would make it such that we would not find meaningful results on the perceived diagnosticity of the initial good deed when the two conditions were averaged together (since participants in the control condition could question why Pat performed the initial good deed, even if they had not been explicitly told nefarious intent was present or absent). Indeed, we did not find a significant main effect of awareness on perceived diagnosticity [$t(97.253) = -0.890$, $p = .376$, $d = -0.312$]. We did, however, find a significant main effect of awareness on overall morality [$t(136) = -4.345$, $p < .001$, $d = -1.521$], trust [$t(108.670) = -3.665$, $p < .001$, $d = -1.256$], and trusting behaviors [$t(130.833) = -3.346$, $p = .001$, $d = -1.076$]. We also found virtually identical results when those who failed the manipulation checks were excluded.

5.2.1. Mediation analyses

Bootstrapped mediation analysis with 5000 bootstrap samples (PROCESS Model 4; Hayes, 2018) also supported the predictions that the differential tendency to discount the diagnosticity of the initial good deed would mediate each of the following dependent variables when tested on its own: overall morality (95% CI [−0.54, −0.06]), trust (95% CI [−0.54, −0.05]), and trusting behaviors (95% CI [−0.64, −0.11]).

5.2.2. Serial mediation analyses

We also tested several forms of serial mediation (PROCESS Model 6; Hayes, 2018). These analyses provided support for a significant indirect effect in which nefarious intent affected the perceived diagnosticity of the initial good deed, which then affected the perceived overall morality of the target, which in turn affected trust in that individual (95% CI [−0.32, −0.04]). We then examined another serial mediation model in which trusting behaviors towards the actor was added to the end of the aforementioned serial mediation model. This analysis revealed a significant indirect effect in which nefarious intent affected perceived diagnosticity, which affected perceived overall morality, which affected trust, and finally affected trusting behaviors (95% CI [−0.14, −0.01]). Finally, we tested an array of alternative causal orderings to assess the viability of serial mediation models beyond what our research predicted and found minimal support for those alternatives (see Appendix B).

5.3. Discussion

The results from Study 3 not only strengthen confidence in the proposed mediator, by revealing that people are more likely to discount the diagnostic value of the initial good deed when it is attributed to nefarious intent than when it is not, but also provide additional support for our predictions. The results also provided no evidence that these effects could be attributed to differences in how observers perceived the bad deed, since this was not affected by the nefarious intent manipulation.

6. Study 4

Study 4 sought to investigate whether these findings would generalize to evaluations of actual behaviors. Hence, our goal was to design a study in which an actor would engage in an actual good deed directed towards an observer followed by an actual bad deed towards the same observer. This was a challenge, given that study participants are generally disinclined to commit clearly bad deeds, especially in the absence of strong situational pressures that could reduce their perceived responsibility for what occurred (Kelley, 1973) and thereby make it harder to make inferences about those participants from their actions. Thus, we settled on a version of the trust game (e.g., Dirks et al., 2011), in which cooperation and defection would represent good and bad deeds, respectively. Because decisions to cooperate and defect in a trust game are likely to seem less meaningful than the kinds of good and bad deeds depicted by the prior studies, this paradigm provides a particularly conservative basis for testing the predictions. However, this design would enable participants to play actor and observer roles and make decisions freely, while also making the initial good deed (cooperation) and eventual bad deed (defection) sufficiently common.

6.1. Method

6.1.1. Participants

Six hundred and forty-six participants currently enrolled as undergraduate students at a private university in the western part of the United States participated in the study in exchange for course credit. We tested this large number of participants because we anticipated that only a subset of these pairs would meet our trust game criteria, in which the observer cooperated throughout the entire game, while the actor initially cooperated (i.e., initial good deed) and later defected (i.e., eventual bad deed). Indeed, by restricting our sample to pairs that met these criteria, we were left with a smaller sample of 126 participants for analysis.[17] We also excluded pairs in which either of the partners failed to correctly answer one or more of the attention check questions. Thus, the final sample consisted of 108 participants (47 female, 61 male). On average, participants were 20.70 years old (ranging from 18 to 30, $SD = 1.80$). We performed sensitivity power analyses in G*Power 3.1, assuming a two-tailed test and an alpha of 0.05. A sample size of $N = 108$ would provide 80% power to detect a main effect of awareness of Cohen's $f = 0.136$, a main effect of perspective of Cohen's $f = 0.236$, and an interaction effect of Cohen's $f = 0.272$.

6.1.2. Research design

We employed a 2 (awareness of subsequent deed: yes vs. no) by 2 (perspective: actor vs. observer) mixed design, with awareness of subsequent deed as a within-subjects factor, and participants randomly assigned to the perspective of either actor or observer.

We designed an online program through which participants could play the trust game with their counterpart. When participants arrived in the research lab, they were seated in individual cubicles each equipped with a computer. Once an even number of participants logged into the game, the program randomly paired them and randomly assigned one partner to the role of actor (Player B) and the other to the role of observer (Player A).[18] The identities of participants' counterparts remained anonymous throughout the study.[19] Once paired, participants were given 5 min to chat via messaging to get to know their counterpart but were told not to disclose any information that would reveal their identities.

Following the 5-min chat, participants were provided the following instructions for the trust game, after which they began playing the game with their counterpart: "You will be randomly assigned to the role of **PLAYER A** or **PLAYER B** to play an Investment Game. For this game you will have the chance to earn tickets. Each ticket represents a chance at a lottery for a $100 Amazon gift card. **PLAYER A** will receive 10 tickets. **PLAYER A** has a choice to either keep all 10 tickets, or to invest them. If **PLAYER A *keeps*** the tickets **PLAYER A** will earn 10 tickets and the game will be over for both players. (Once the game is over, players will be directed to complete the post-questionnaire). If **PLAYER A *invests*** the tickets, the tickets will be multiplied by 4, resulting in 40 tickets which are all transferred to **PLAYER B**. **PLAYER B** then decides whether to keep the tickets or to send half back to **PLAYER A**. If **PLAYER B *keeps*** the tickets, then **PLAYER B** earns 40 tickets, **PLAYER A** earns 0 tickets, and the game is over for both players. If **PLAYER B *shares*** the tickets, then these tickets are split evenly (**PLAYER B** earns 20 tickets, **PLAYER A** earns 20 tickets), and the game is played for another round."

We were only interested in the pairs in which actors cooperated in round #1 by sharing half of the tickets with the observer (i.e., the initial good deed), then defected in a later round by keeping all of the tickets for him/herself (i.e., the eventual bad deed). If the actor cooperated in round #1, both players were then asked to evaluate this initial good deed by the actor. Partners then played additional rounds until one of them defected or the game automatically ended after five rounds. When the actor eventually defected in a later round, both players were once again asked to evaluate the actor's initial good deed from round #1. At no point were participants able to see their partner's responses. If the observer defected first or neither partner defected through the final round, participants were given filler questions to answer instead.

[17] We stated in our preregistration (As Predicted #30022, https://aspredicted.org/FQX_ZN6) that we would collect data from 124 participants who met our trust game criteria, but we ended up with 126 participants.

[18] If an odd number of students showed up for the study session, one student was randomly selected, given credit for participating, and dismissed.

[19] When there were not enough participants in the study session to maintain the anonymity of partners, all study participants were given participation credit and dismissed.

6.1.3. Measures[20]

Participants responded to attention check questions after the actor defected (thus ending the game). The attention check questions asked how many rounds they had just played and which player ended the game. Responses were checked against the data from the game. Participants responded to each of the multi-item scales from the prior studies twice, immediately after the initial good deed (g) and immediately after the subsequent bad deed (b): perceived diagnosticity ($\alpha_g = 0.92$, $\alpha_b = 0.96$), nefarious intent ($\alpha_g = 0.74$, $\alpha_b = 0.88$), perceived morality ($\alpha_g = 0.84$, $\alpha_b = 0.85$), and trust ($\alpha_g = 0.50$, $\alpha_b = 0.75$). Items were adapted to reflect the actions of the game [e.g., "Your counterpart's action (sharing the tickets in the 1st round of the game played) is reflective of his/her moral character"]. We also developed a single item to assess trusting behavior towards the actor. Participants were informed they might have a chance to play one more round of the investment game with the same counterpart and asked whether they (in the Observer condition) or their counterpart (in the Actor condition) would invest all the tickets in that next round, on a 7-point scale (1 = very unlikely, 7 = very likely).

6.2. Results

Correlations for the study variables are provided in Table 7. Means and standard deviations by condition are provided in Table 8. Because participants were paired up for the game and each actor and observer within a pair was meaningfully yoked to his/her counterpart, we treated the variable of perspective as a repeated measures factor in order to account for potential intra-dyad correlations (Warner, 2013) and conducted two-way repeated measures ANCOVAs. Since partners could play between two to five rounds before the actor defected, we controlled for the number of rounds played. We also ran ANOVAs (i.e., without controlling for number of rounds played) and found similar results, with only one exception to be mentioned later. And though this analytical approach was preregistered for this study and it provided the most straightforward way to report the findings, we also ran tests via multi-level modeling (MLM)[21] and found that the results were generally robust across approaches (with exceptions mentioned in later footnotes).

Although awareness of the subsequent bad deed did not exert a significant main effect on perceived diagnosticity [$F(1, 52) = 2.495$, $p = .120$, $\eta_p^2 = 0.046$], it did exert significant main effects on nefarious intent [$F(1, 52) = 22.164$, $p < .001$, $\eta_p^2 = 0.299$], overall morality [$F(1, 52) = 38.548$, $p < .001$, $\eta_p^2 = 0.426$], trust [$F(1, 52) = 14.196$, $p < .001$, $\eta_p^2 = 0.214$], and trusting behavior [$F(1, 52) = 17.100$, $p < .001$, $\eta_p^2 = 0.247$]. Participants imputed greater nefarious intent, perceived lower overall morality, and exhibited lower trust and trusting behavior towards the actor after the actor's later bad deed than before the bad deed.[22] We did not find significant main effects of perspective on any of our measures with the exception of trust [$F(1, 52) = 5.954$, $p = .018$, $\eta_p^2 = 0.103$]. Observers rated the actor as significantly less trustworthy ($M = 3.63$, $SD = 0.80$) than actors rated themselves ($M = 4.50$, $SD = 0.76$).

However, these results were ultimately qualified by significant awareness x perspective interaction effects (See Fig. 5). Although we did not find a significant awareness x perspective interaction effect on perceived diagnosticity [$F(1, 52) = 0.848$, $p = .361$, $\eta_p^2 = 0.016$], we did find significant interaction effects for nefarious intent [$F(1, 52) = 10.272$, $p = .002$, $\eta_p^2 = 0.165$], overall morality [$F(1, 52) = 11.063$, $p = .002$, $\eta_p^2 = 0.175$], trust [$F(1, 52) = 10.710$, $p = .002$, $\eta_p^2 = 0.171$], and trusting behavior [$F(1, 52) = 9.789$, $p = .003$, $\eta_p^2 = 0.158$]. Awareness of the actor's subsequent bad deed led to a greater imputation of nefarious intent and a greater decrease in perceived morality, trust, and trusting behavior for observers than actors.[23]

6.2.1. Mediation analyses

Although methods for analyzing mediation, moderation, and moderated mediation for between-subjects designs have been well established, there was no clearly established method for testing these effects in repeated-measures designs involving a moderator (i.e., perspective) that may not be considered purely independent due to clustering at the dyadic level (Montoya, 2018; Montoya, 2019; Montoya & Hayes, 2017). We therefore considered several existing approaches and ultimately chose one based on Valente and MacKinnon's Difference Score Model (2017).[24] This approach involved thinking about our moderator (i.e., perspective) as our independent variable and using difference scores to capture change over time for each of our dependent variables (Montoya, 2018; Valente & MacKinnon, 2017). We recognize the use of difference scores alters the interpretation of the mediation models, but we felt this approach was still theoretically meaningful for our purposes since we were interested in how assessments of the initial good deed changed in light of the subsequent bad deed.

Our only deviation from Valente and MacKinnon (2017) was that although their data involved a between-subjects independent variable, we continued to treat perspective as a repeated-measures variable as we previously did in our ANCOVAs, since actors and observers were clustered within dyads in our study. Also, since MEMORE does not allow for covariates, we were unable to control for the number of rounds played. However, we found that not controlling for rounds produced largely similar results with only one exception in which the interaction effect on

[20] We also measured participants' judgments of the goodness of the cooperation itself with a two-item scale that was tailored to this study (e.g., "to what extent do you think your counterpart's action (sharing the tickets in the 1st round of the game played) was good?"). This was to verify the conclusion from our early exploratory efforts that those judgments of the act itself would be less effective at explaining the kinds of influences we sought to investigate than perceived diagnosticity in this behavioral trust game context. As we found in those early exploratory efforts, although the results for participants' judgments of the act were consistent with what we found for perceived diagnosticity, those judgments of the act itself were not found to play a significant role in any of the mediation models we had tested. These findings provide further support for the decision made based on our initial exploratory efforts to focus on perceived diagnosticity rather than judgments of the act itself.

[21] The MLM focused on two-level models with time (or instances of evaluation) at level 1, nested within individuals at level 2. Given the relatively low intraclass correlation coefficients (ICCs) at the dyadic level (<0.10) and because we were not interested in effects at the dyadic level, we simply accounted for potential intra-dyad correlations by specifying clustered standard errors at the dyadic level (Cameron & Miller, 2015).

[22] When we ran the analyses via MLM, we did find a significant main effect of awareness on diagnosticity ($b = -0.66$, robust $SE = 0.16$, $p < .001$).

[23] When we ran the analyses via MLM, we did find a marginally significant awareness x perspective cross-level interaction effect on perceived diagnosticity ($b = 0.46$, robust $SE = 0.25$, $p = .066$), but we did not find a significant interaction effect on trusting behavior ($b = 0.57$, robust $SE = 0.44$, $p = .191$). When we did not control for number of rounds played in our MANOVA, we did not find a significant interaction effect on trusting behavior, $F(1, 53) = 1.711$, $p = .196$, $\eta_p^2 = 0.031$. Although these supplementary analyses did not offer further support for a direct interaction effect on trusting behavior, we did find through a simple regression that change in trust significantly predicted change in trusting behavior [$b = 0.62$, $t(106) = 3.211$, $p = .002$]. Furthermore, our mediation analyses showed that perspective had a significant indirect effect on change in trusting behavior through change in trust (as will be discussed under mediation analyses).

[24] As in our main analyses, we explored the option of treating our dataset as a multi-level dataset with time at level 1, nested within individuals at level 2, while controlling for clustering at the dyadic level. Nicholas Rockwood's MLmed macro would enable us to run moderated mediation with multilevel data up to two levels (Rockwood, 2017). However, it would not allow us to account for clustering at the dyadic level, given that MLmed does not have an option to either specify clustered standard errors or include a third level.

Table 7
Correlations for Study Variables (Study 4).[a, b]

Variable	1	2	3	4	5	6	7	8	9	10
1. Nefarious Intent (t1) [a]	–									
2. Nefarious Intent (t2) [b]	0.546***	–								
3. Diagnosticity (t1) [a]	−0.125	0.071	–							
4. Diagnosticity (t2) [b]	0.166	0.076	0.642***	–						
5. Overall Morality (t1) [a]	−0.168	−0.038	0.613***	0.472***	–					
6. Overall Morality (t2) [b]	0.060	−0.344***	0.301**	0.395***	0.420***	–				
7. Trust (t1) [a]	−0.183	−0.197*	0.338***	0.340***	0.503***	0.351***	–			
8. Trust (t2) [b]	−0.033	−0.458***	0.213*	0.320***	0.172	0.677***	0.456***	–		
9. Trusting Behaviors (t1) [a]	−0.103	0.103	0.107	0.036	0.022	−0.080	0.002	−0.150	–	
10. Trusting Behaviors (t2) [b]	−0.162	−0.289**	0.120	0.163	0.020	0.257**	−0.032	0.194*	0.103	–

* $p < .05$. ** $p < .01$. *** $p < .001$.
[a] t1 = after initial good deed.
[b] t2 = after subsequent bad deed.

Table 8
Means and Standard Deviations by Awareness of Subsequent Deed x Perspective (Study 4).

			Awareness of Subsequent Bad Deed		
			No *M (SD)*	Yes *M (SD)*	Total *M (SD)*
Perspective	Observer *M (SD)*	Nefarious Intent	3.56 (1.15)	4.92 (1.49)	4.24 (1.14)
		Diagnosticity	4.48 (1.33)	3.81 (1.35)	4.15 (1.19)
		Overall Morality	5.34 (0.76)	3.94 (0.86)	4.64 (0.63)
		Trust	4.34 (0.83)	2.93 (1.09)	3.63 (0.80)
		Trusting Behavior	6.15 (1.30)	3.41 (2.08)	4.78 (1.17)
	Actor *M (SD)*	Nefarious Intent	4.09 (1.18)	4.44 (1.36)	4.26 (1.19)
		Diagnosticity	4.60 (1.52)	4.40 (1.58)	4.50 (1.43)
		Overall Morality	5.24 (0.93)	4.77 (0.92)	5.00 (0.86)
		Trust	4.75 (0.94)	4.25 (0.89)	4.50 (0.78)
		Trusting Behavior	5.31 (1.49)	3.15 (1.95)	4.23 (1.37)
	Total *M (SD)*	Nefarious Intent	3.83 (1.19)	4.68 (1.44)	4.25 (1.16)
		Diagnosticity	4.54 (1.42)	4.10 (1.49)	4.32 (1.32)
		Overall Morality	5.29 (0.85)	4.36 (0.98)	4.82 (0.77)
		Trust	4.55 (0.90)	3.59 (1.20)	4.07 (0.90)
		Trusting Behavior	5.73 (1.45)	3.28 (2.01)	4.50 (1.30)

trusting behavior was no longer significant when we removed rounds as a covariate, $F(1, 53) = 1.711$, $p = .196$, $\eta_p^2 = 0.031$. Thus, running mediation analyses without controlling for rounds represents a more conservative approach to testing our predictions.

Bootstrapped mediation analyses with 5000 samples (MEMORE Model 1) supported the prediction that the tendency to discount the diagnosticity of the initial good deed would be mediated by the extent to which people impute nefarious intent onto the actor's initial good deed (95% CI [0.04, 0.80]). However, given the lack of support for perceived diagnosticity (and the subsequent lack of support for our mediation models involving change in perceived diagnosticity as a mediator), we instead ran mediation analyses that focused on examining how our manipulations affected the predicted causal sequence that followed the perceived diagnosticity measure.

Bootstrapped mediation analysis with 5000 bootstrap samples (MEMORE Model 1) supported a significant indirect effect of perspective on change in trust through change in overall morality (95% CI [0.33, 0.99]). We also tested a serial mediation model that added change in trusting behavior to the end of the previous model. The bootstrapped serial mediation analysis with 5000 bootstrap samples (MEMORE Model 1) did not provide support for a significant indirect effect of perspective on change in trusting behavior through change in overall morality and change in trust (95% CI [−0.16, 0.88]). Next, to examine whether changes in trust would affect changes in trusting behavior, we ran a mediation model that tested whether perspective would affect change in trusting behavior through change in trust. The bootstrapped mediation analysis with 5000 bootstrap samples (MEMORE Model 1) provided support for a significant indirect effect of perspective on change in trusting behavior through change in trust (95% CI [0.08, 1.24]). Finally, we tested alternative causal orderings to assess the viability of serial mediation models beyond what our research predicted and found no support for those alternatives (see Appendix B).

6.3. Discussion

These results broadly generalize support for our predictions to actors and observers engaging in and witnessing actual behaviors. Awareness of an actor's subsequent bad deed led people to impute nefarious intent onto the actor's initial good deed, and more importantly, observers did so to a greater extent than actors themselves. Actors and observers also differed in the extent to which their evaluations of overall morality, trust, and trusting behavior changed in light of the actor's subsequent bad deed. However, we did not find direct effects on perceived diagnosticity.

Although the lack of support for perceived diagnosticity meant that we generally did not find support for mediation models that used this variable as a mediator, we largely found support for our predicted mediation chain otherwise. Most importantly, we found that awareness of the subsequent bad deed affected perceived diagnosticity in an indirect manner through actors' and observers' differential imputation of nefarious intent. This is notable, since this supports the primary theoretical model that represents the core set of predicted relationships for this research. We also found that differences in actors' and observers' perceptions of the actor's overall morality mediated the degree of change in trust in the actor and that changes in trust towards the actor mediated changes in people's (expected) likelihood of engaging in trusting behavior. Thus, even though we did not find direct support for perceived diagnosticity, we did find support for mediation both up to and following that measure.

We believe the lack of direct effects for perceived diagnosticity can be attributed to limitations in operationalizing the good and bad deeds in this study, given the challenges of identifying a paradigm in which actors would willingly engage in more meaningful bad deeds. The good and bad deeds in Study 4 were thus much more impoverished in that they entailed cooperating and defecting in the context of a game, actions

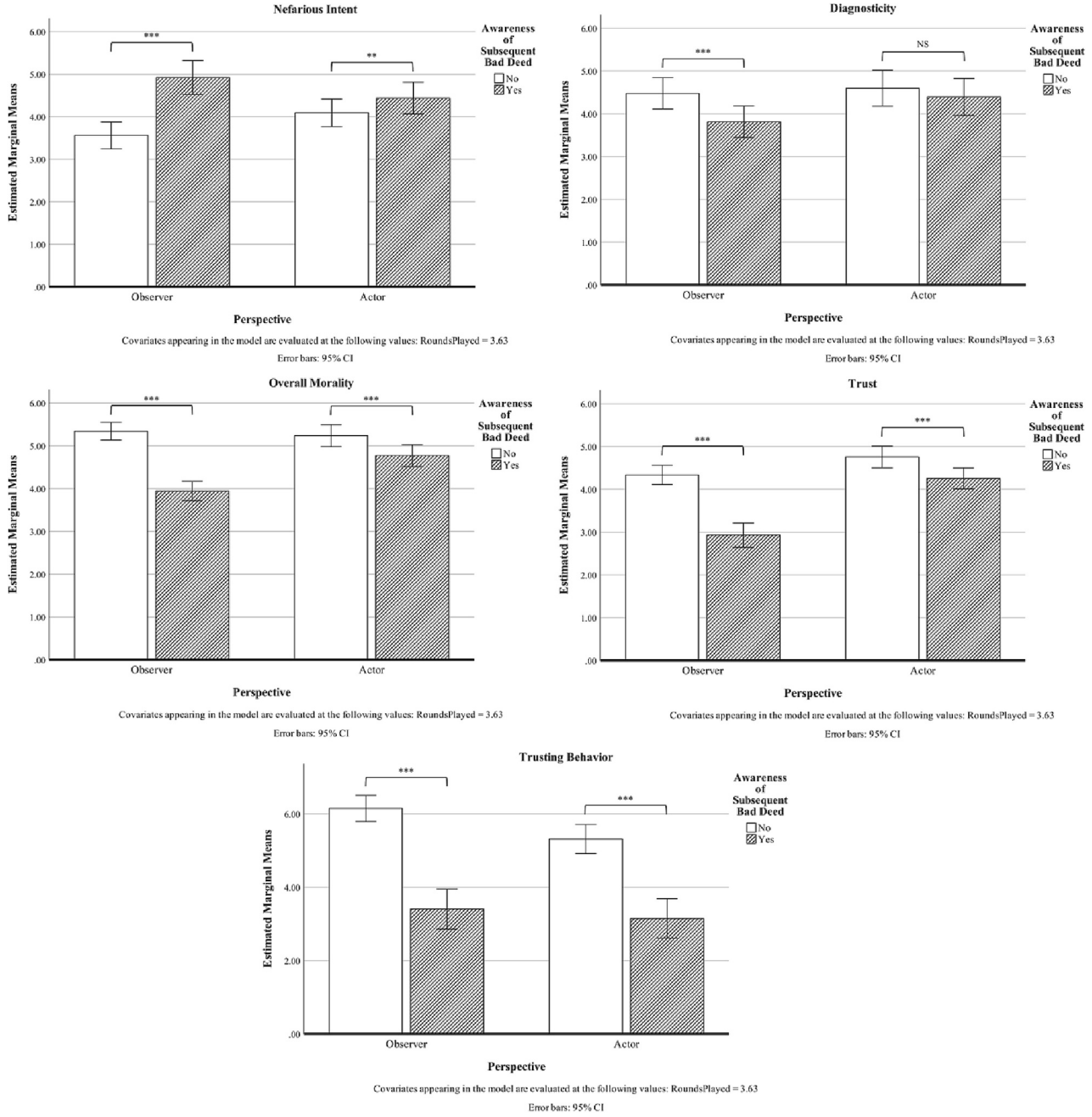

Fig. 5. Means by Awareness of Subsequent Deed x Perspective (Study 4).
** $p < .01$. *** $p < .001$. NS = Not Significant.

participants may have been less inclined to perceive as ethical or unethical, particularly if they held normative expectations about early cooperation and later defection in this investment game context. Thus, given these limitations, Study 4 should be considered a particularly conservative test of our predictions, and it is noteworthy that its results still supported the core set of relationships in our primary theoretical model, as well as our theoretical reasoning for all of the other measures.

7. Study 5

Finally, given the constraints Study 4 faced operationalizing the good and bad deeds (so that actors would willingly engage in those behaviors), we conducted a second behavioral experiment that would allow us to operationalize the good and bad deeds more robustly, directly manipulate the number of interactions, and also assess trusting behaviors with different measures. Thus, we designed a study in which a computer confederate would play the actor role and focused on how observers would respond to that confederate's actions.

7.1. Method

7.1.1. Participants

Two hundred and forty participants from the United States were recruited through Prolific and were each paid $3.00 to complete the study as well as a bonus payment (ranging from $0.05 to $0.44) for the amount each participant earned through the study. Although 12

participants suspected that they were interacting with a computer rather than another participant, excluding them did not affect support for our predictions. We, therefore, retained the total sample of 240 participants (130 female, 106 male, 4 other).[25] On average, participants were 34.65 years old (ranging from 18 to 77, $SD = 11.87$). We performed a sensitivity power analysis in G*Power 3.1 testing our main effects and interaction effects with a MANOVA, assuming a two-tailed test and an alpha of 0.05. A sample size of $N = 240$ would provide 80% power to detect an effect of Cohen's $f = 0.192$. We also performed a sensitivity power analysis testing our main effects with an ANCOVA; a sample size of $N = 240$ would provide 80% power to detect an effect of Cohen's $f = 0.182$.

7.1.2. Research design

We randomly assigned participants to one of four experimental conditions in a 2 (awareness of subsequent bad deed: yes vs. no) by 2 (total number of rounds: two vs. four) between-subjects design.

We developed an online game in which participants were informed that they would be paired with an anonymous counterpart for the game and that players would be randomly assigned to the role of Decider or Advisor.[26] In actuality, all participants were assigned to the Decider role and would interact with a computer-programmed 'Advisor'. Once assigned to the Decider role, participants were given the following instructions: "During the game, you and your counterpart will have the chance to earn extra money (in addition to your study participation payment). In the game, two possible monetary payments will be available to you and your counterpart, Option A and Option B. For example: Option A: Decider gets **X**¢, Advisor gets **Y**¢ / Option B: Decider gets **Z**¢, Advisor gets **W**¢. First, only your counterpart, the **Advisor**, will be able to see the payments associated with Option A and Option B. You will **not** be able to see the payment amounts. Your counterpart (the Advisor) will be instructed to send you a message recommending one of the options for you to choose. You, the **Decider**, will then be given the task of choosing between Option A and Option B. You do **not** have to follow the advice from the Advisor. After you make a choice, you will be able to see the payments associated with each of the options. The extra money that you and your counterpart each receive will depend on the option you choose."

Payment options varied for each round, and participants were informed that at the end of the game, each participant would be sent a bonus payment in the amount they had accumulated throughout the rounds. Although participants were informed that the task would be repeated for a randomly determined number of rounds, we programmed the game to end after two or four rounds, depending on the condition. Earlier rounds always involved the Advisor engaging in a good deed by honestly advising the participant to choose the option that would earn the participant the most money, even at a loss to the Advisor. We varied whether participants would be made aware of a bad deed in the final round of the game by having the Advisor either engage in another good deed or engage in a bad deed. In the bad deed, the Advisor would advise the participant to choose the option that would earn the participant the least amount, while gaining the Advisor more. Once the game ended, participants would be asked to evaluate their counterpart's (i.e., the Advisor's) good deed from the first round of the game.

7.1.3. Measures

We assessed the perceived diagnosticity of the initial good deed ($\alpha = 0.97$), nefarious intent ($\alpha = 0.93$), overall morality ($\alpha = 0.98$), and trust ($\alpha = 0.93$) using the same multi-item scales as the prior studies. Items were adapted to reflect the actions of the game [e.g., "Your counterpart's earlier action (advising you to choose the option that would earn you the most money in Round 1) is reflective of his/her moral character"].

We also developed two new items to assess trusting behaviors towards the actor. Participants were asked, if given the chance to play another round of the game with the same counterpart, how willing they would be to play again (1 = very unwilling, 7 = very willing). Participants were also asked, if given the chance to play another round of the game, whether they would choose to play with the same counterpart or a new counterpart (1 = strongly prefer new counterpart, 7 = strongly prefer same counterpart). Responses on these two items were averaged to create a composite measure of behavioral trust ($r_{SB} = 0.83$).

We assessed negative affect by asking participants to indicate how they felt using the following adjectives: negative, concerned, bothered, frustrated, tense, anxious, and distressed (1 = does not apply at all right now, 7 = applies very much right now; Wakslak, Jost, Tyler, & Chen, 2007). Responses on these seven items were averaged to compose a measure of negative affect ($\alpha = 0.94$).

Finally, we included an open-ended question asking participants to tell us what they thought about their counterpart, in order to gauge whether participants suspected they were interacting with a computer rather than another participant.

7.2. Results

Correlations for the study variables are provided in Table 9. Means and standard deviations by condition are provided in Table 10. Preliminary two-way ANOVAs did not reveal a significant awareness x rounds interaction effect on Diagnosticity [$F(1, 236) = 0.464$, $p = .496$, $\eta_p^2 = 0.002$], consistent with our expectation that the number of rounds would not affect discounting of the diagnosticity of the initial good deed. However, because these analyses revealed significant main effects of rounds played on four of our dependent measures, as well as significant awareness x rounds interactions on two dependent measures, we treated number of rounds as a covariate in our main analyses.[27] We also controlled for negative affect in those analyses, to rule out the possibility that our effects could be explained by negative affect influencing participants' recall or interpretation of past events.

One-way ANCOVAs controlling for rounds and negative affect revealed significant main effects of awareness on nefarious intent [$F(1, 236) = 153.169$, $p < .001$, $\eta_p^2 = 0.394$], diagnosticity, [$F(1, 236) = 44.398$, $p < .001$, $\eta_p^2 = 0.158$], overall morality [$F(1, 236) = 202.974$, $p < .001$, $\eta_p^2 = 0.462$], trust [$F(1, 236) = 264.407$, $p < .001$, $\eta_p^2 = 0.528$], and trusting behaviors [$F(1, 236) = 129.787$, $p < .001$, $\eta_p^2 = 0.355$]. Participants who evaluated the actor's initial good deed after witnessing the actor's subsequent bad deed imputed greater nefarious intent onto the initial good deed, perceived the initial good deed as less diagnostic of the actor's moral character, evaluated the actor as lower in overall morality, reported lower trust, and were less willing to engage in trusting behaviors towards the actor compared to those who did not witness the subsequent bad deed (See Fig. 6). Supplementary analyses showed that our results were entirely consistent even when we did not control for number of rounds or negative affect.

7.2.1. Mediation analyses

We conducted bootstrapped mediation analyses with 5000 bootstrap

[25] As Predicted #98930, https://aspredicted.org/PWY_X37

[26] This game design was a multi-round Deception Game (Gneezy, 2005) modified for online exchange with a computer-programmed confederate.

[27] We found significant main effects of rounds on Diagnosticity [$F(1, 236) = 4.053$, $p = .045$, $\eta_p^2 = 0.017$], overall morality [$F(1, 236) = 14.909$, $p < .001$, $\eta_p^2 = 0.059$], trust [$F(1, 236) = 12.487$, $p < .001$, $\eta_p^2 = 0.050$], and trusting behavior [$F(1, 236) = 7.622$, $p = .006$, $\eta_p^2 = 0.031$]. We found significant interaction effects on overall morality [$F(1, 236) = 4.561$, $p = .034$, $\eta_p^2 = 0.019$], and trust [$F(1, 236) = 3.865$, $p = .050$, $\eta_p^2 = 0.016$].

Table 9
Correlations for Study Variables (Study 5).

Variable	1	2	3	4	5
1. Nefarious Intent	–				
2. Diagnosticity	−0.616***	–			
3. Overall Morality	−0.828***	0.665***	–		
4. Trust	−0.792***	0.602***	0.899***	–	
5. Trusting Behaviors	−0.640***	0.470***	0.758***	0.749***	–

*** $p < .001$.

Table 10
Means and Standard Deviations by Awareness of Subsequent Deed (Study 5).

	Awareness of Subsequent Bad Deed		
	No *M (SD)*	Yes *M (SD)*	Total *M (SD)*
Nefarious Intent	2.30 (0.98)	4.72 (1.37)	3.50 (1.70)
Diagnosticity	5.89 (1.02)	4.58 (1.48)	5.24 (1.42)
Overall Morality	6.05 (0.84)	3.68 (1.21)	4.88 (1.58)
Trust	5.52 (0.95)	2.80 (1.18)	4.17 (1.73)
Trusting Behaviors	6.54 (0.72)	4.27 (1.61)	5.41 (1.68)

samples, controlling for number of rounds and negative affect (PROCESS Model 4; Hayes, 2018).[28] These analyses supported the prediction that the tendency to discount the diagnosticity of the initial good deed would be mediated by the imputation of nefarious intent (95% CI [0.70, 1.44]), as well as the predictions that this discounting of the diagnosticity of the initial good deed would mediate each of the following dependent variables when tested on its own: overall morality (95% CI [0.31, 0.77]), trust (95% CI [0.25, 0.65]), and trusting behaviors (95% CI [0.08, 0.46]).

7.2.2. Serial mediation analyses

We also tested several serial mediation models, controlling for number of rounds and negative affect (PROCESS Model 6; Hayes, 2018). These analyses provided support for a significant indirect effect in which nefarious intent affected the perceived diagnosticity of the initial good deed, which then affected the perceived overall morality of the target (95% CI [0.13, 0.48]). We also found support for a significant indirect effect in which nefarious intent affected perceived diagnosticity, which then affected the perceived overall morality of the target, which in turn affected trust in the target individual (95% CI [0.09, 0.34]). Furthermore, these analyses provided support for a significant indirect effect in which nefarious intent affected perceived diagnosticity, which then affected perceived overall morality, which affected trust, and finally affected trusting behaviors (95% CI [0.01, 0.11]). Finally, although we found support for some alternative serial mediation models, we did not find consistent support for those alternatives across our studies (see Appendix B).

7.3. Discussion

The findings from this study further extend support for our predictions to the context of a different behavioral experiment. Moreover, this study's ability to operationalize its good and bad deeds in a more robust manner allowed us to obtain not only indirect support, but also strong and consistent direct support for the role of perceived diagnosticity in these predicted relationships.

8. General discussion

The purpose of this paper was to investigate a critical blind spot in reputation management. Although past research has offered important insights into the ways in which people attempt to manage their moral standing, that work has typically presumed that such appraisals would involve the aggregation of essentially static interpretations of a target's discrete acts. Our research reveals, however, that such interpretations are far from static, and that they can change far more than targets realize as new events unfold.

We found this retrospective effect is not only more likely to occur when a prior good deed is followed by a bad deed than when this order was reversed, but that it also occurred to a greater degree with observers than actors. This is because observers were more inclined than actors to reinterpret the prior good deed as an attempt to set up others for the bad deed they had planned all along (i.e., to infer nefarious intent), and in turn make lower assessments of the target's morality and trustworthiness. These predicted effects were supported in different relational and study contexts, with different types of study populations, and regardless of how many good deeds occurred before the bad deed was committed. Finally, although the correlational nature of the mediation analyses prevents us from definitely ruling out alternative mediation sequences, Study 3's direct manipulation of the focal mediator (nefarious intent), as well as the fact that far more consistent support was found for the predicted serial mediation sequence than any alternative sequence, helps lend at least some credence to our theoretical model.

These findings cannot be explained by basic mechanisms like self-serving motivations, actors simply being less likely to believe that they have nefarious intent or low morality than their observers, inherent differences in information diagnosticity, or the notion that *any* piece of information may become less diagnostic as more information becomes available. Self-serving motivations, for example, would suggest that actors are likely to interpret the implications of their own behavior more positively than their observers. However, we found that actors and observers did not differ in their interpretations of the bad deed (a deed that actors should be particularly motivated to interpret in self-serving ways) and that observers viewed the actors' good deed even more positively than the actors themselves, when these acts were considered on their own (Pretest #3). Likewise, past research on people's tendency to form rosier moral views of themselves than others has found that people are nevertheless relatively accurate when asked to judge how others would view them (Rom & Conway, 2018), indicating that people would be at least somewhat aware of their self-serving tendencies. Yet our findings were just as strong regardless of whether actors judged themselves (Study 1) or how others would judge them (Study 2), suggesting that our findings are ultimately beyond people's awareness.

Similarly, support for the notion that the findings stem from actors simply being less likely to believe they would have nefarious intent or low morality than those who observe them seems mixed at best. That possibility suggests that actors would attribute less nefarious intent for the initial good deed (and believe actors would have greater overall morality) than their observers *regardless* of participants' awareness of the subsequent bad deed (i.e., since even in cases where the subsequent bad deed was not revealed, the nefarious intent measure asks participants to gauge the extent to which the initial good deed was intended to set the stage for a subsequent bad deed). And though we do find such main effects in Study 1, those effects were ultimately qualified by significant interactions with participants' awareness of the subsequent bad deed, which reveal that this actor vs. observer difference primarily occurred when the subsequent bad deed became known. Moreover, neither Studies 2 nor 4 found support for these main effects of perspective and instead, once again, provided robust support for our predicted interactions.

Finally, although a) differences in information diagnosticity might

[28] Supplementary analyses found virtually identical results when we did not control for number of rounds or negative affect in our mediation analyses. We also conducted moderation mediation analyses with number of rounds as a moderator rather than a covariate. Consistent with our expectation that number of rounds would not affecting discounting, we did not find significant indices of moderated mediation but did find significant conditional indirect effects for all of our models. Full results are available upon request.

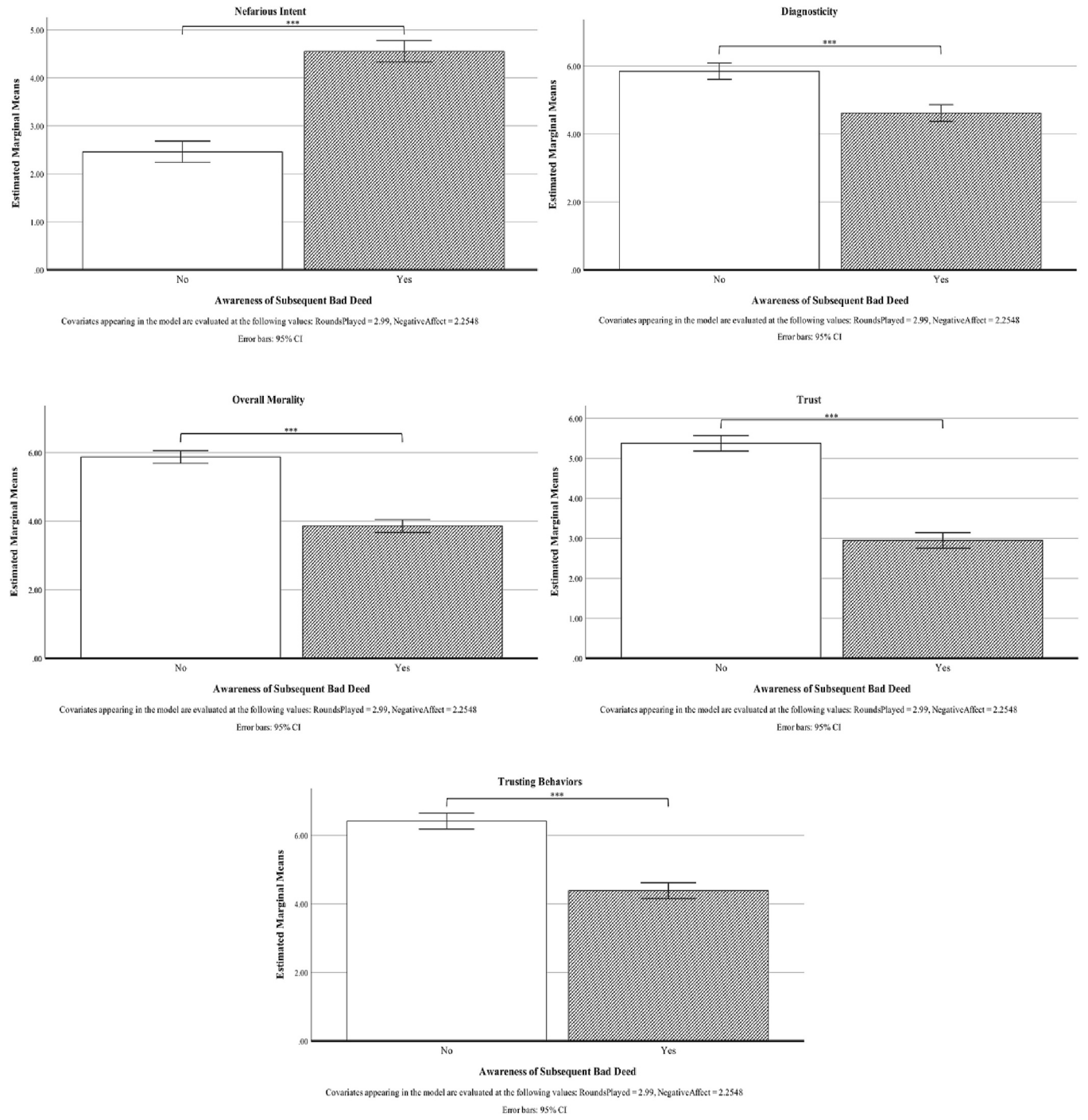

Fig. 6. Means by Awareness of Subsequent Deed (Study 5).
*** $p < .001$.

explain why bad deeds are considered more significant than good deeds in general, and b) the notion that any prior act becomes less diagnostic as additional behaviors are considered might explain why the perceived diagnosticity of any deed might generally decline as new information is added (Anderson, 1962), it is unclear how those mechanisms could account for why observers would discount the diagnosticity of the initial good deed more than actors once the subsequent bad deed became known. To do so, observers should have perceived the bad deed to be more diagnostic than actors so that harsher observer assessment would counterbalance observers' initial positive views (based on the initial good deed alone) to a greater degree when both the good deed and bad deed are combined. Yet our findings reveal that observers did not differ from actors in how diagnostic they perceived the bad deed to be (and, if anything, considered the good deed to be even more diagnostic than actors) when each deed was considered on its own. Hence, these kinds of diagnosticity considerations offer little reason why observers would suddenly discount the initial good deed far more than actors once the subsequent bad deed was revealed.

Of course, these observations are not meant to rule out every possible way in which these mechanisms might have exerted some influence. For example, our research drew on the potential implications of information diagnosticity to explain why the predicted retrospective mechanism would be limited to prior good deeds. However, the broader set of findings suggest that actors' and observers' differential tendencies to

reassess past behavior, when subsequent acts occur, may be more meaningfully understood from the standpoint of how people manage this ethical accounting process over time, rather than as a more generic feature of individual cognition. And this ultimately offers important implications for research on how one's moral standing might be managed.

8.1. Theoretical implications

Although research on moral judgment has generally been grounded on the notion that the implications of one's past deeds would remain as a repository of inferences to which new ones are aggregated (e.g., Effron and Monin, 2010; Mazar et al., 2008; Nisan, 1990), other findings have suggested that those implications may change as new events unfold. Yet the literature has offered little evidence to directly support the latter possibility in the moral domain. Moreover, the research that does exist at least implicitly suggests this kind of retrospective updating would be limited to when people behave ethically and unethically with respect to the exact same kind of deed (i.e., by engaging in blatant forms of hypocrisy) rather than apply more broadly to cases where the good and bad deeds might differ (e.g., Jordan et al., 2017). Hence, little has been done to compare these competing possibilities, let alone reconcile them into an integrative account of what should happen (Kim et al., 2018). Our findings, thus, contribute to this literature by: 1) providing more direct support for the notion that people can revise their interpretations of what occurred in the past, 2) demonstrating that this tendency can generalize well beyond cases where people act ethically and unethically with respect to the exact same kind of deed, and 3) ultimately providing much needed insight into *when* and *why* that retrospective updating might occur.

Indeed, the premise that the implications of one's past acts would remain as a repository against which subsequent deeds might add or subtract was still supported in many cases. Whether the implications of one's past acts would a) remain static or b) change as new events unfold ultimately depended on both the sequence of events and perspective. This highlights the need for more integrative research that moves beyond narrower efforts to support each possibility on its own, to seek deeper insight into the conditions where each of these different intertemporal processes is more likely to be supported (Kim et al., 2018).

Moreover, by revealing how this ethical accounting process ultimately affects trust, our findings also shed light on the dynamic process through which perceptions that are related to morality, such as trustworthiness, can arise and diminish. In particular, the trust literature has observed that even a high level of trust built over a substantial period of time can be destroyed even by a single transgression (e.g., Kim, Dirks, & Cooper, 2009). However, if this high trust is based on a lengthy period of trustworthy behavior lasting months or even years, the notion that just one untrustworthy act would outweigh all that countervailing evidence may seem less convincing. Our findings, however, may help address this apparent explanatory shortcoming, by suggesting that the potential for one untrustworthy act to be so devastating may arise not only because it can be weighed more heavily than trustworthy behavior, but also because it can cause that prior trustworthy behavior to be reinterpreted as nefarious attempts to set the stage for the transgression that had been planned all along.

Further, the fact that such influences ultimately led actors and observers to differ so markedly in their character assessments may ultimately help explain why trust may be so difficult to repair after a violation (e.g., Bottom, Gibson, Daniels, & Murnighan, 2002; Harmon, Kim, & Mayer, 2015; Kim et al., 2017). In particular, past research in this domain has framed the trust repair process as a matter of resolving discrepant beliefs about the target, with perceivers believing trust is not warranted and the target believing that greater trust would be deserved (Kim et al., 2009). However, this literature has not delved deeply into why this might be the case. The present research sheds light on this issue by revealing how perceivers and targets may evaluate the target's trustworthiness quite differently, in the aftermath of a transgression, due to their differential tendencies to respond to that transgression by reevaluating acts from the past. By doing so, this research underscores how transgressors may fail to appreciate not only how their pre-transgression behaviors would be reinterpreted, but also what this would entail for how their moral character and trustworthiness would be assessed. And this can thereby help explain why their subsequent attempts to address such incidents, and ultimately repair trust, may so often fall short.

8.2. Practical implications and future directions

The findings from this research also highlight important practical implications. Most notably, the results reveal that people may be far less capable of maintaining their moral standing than they think, at least in the eyes of others, and that this can lead them to become morally bankrupt even when this is something they have deliberately sought to avoid. This shortcoming arises from the fact that people's assumption that their prior good deeds would persist to counterbalance their subsequent bad acts may actually be wrong. And this possibility underscores the notion that people may not be able to give themselves the kinds of allowances to engage in occasional unethical acts they might believe they have earned, given that even one unethical act can transform how their past actions are interpreted.

These findings, in turn, highlight the need for future research to obtain better insight into how and when this ethical accounting process might unfold. One might wonder, for example, whether members of different cultures would engage in this ethical accounting process differently due to their different views about the degree of control people possess relative to the situation (Maddux, Kim, Okumura, & Brett, 2011; Morris, Menon, & Ames, 2001). Studies might, furthermore, investigate whether these effects would differ as a function of the relative power of the actor and observer, differences in group membership, or a host of demographic and social characteristics (Kim et al., 2017). Finally, one might also consider whether these effects might depend on the ambiguity of the good and/or bad deed, what kinds of situational features may lead people to feel more or less able to engage in these re-evaluations, and how these effects might depend on the level of knowledge the evaluator believes he/she has already gained about the actor.

These kinds of considerations ultimately highlight the importance of obtaining greater insight into how people evaluate, and account for, the implications of their ethical and unethical behavior over time. The fact that actors and observers can differ so strikingly in these activities, and its potential to create such marked differences in their assessments, may create major complications for those seeking to maintain their moral standing, reputations, and ability to engage in fruitful social interactions with others. Thus, to the extent that people continue to manage their ethically-relevant behavior in a manner that seems to misjudge how those actions would actually be perceived, concerted attention and efforts to unpack what may shape this process seem long overdue.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2023.104461.

Data availability

Data will be made available on request.

Appendix A. Moderated mediation models (Studies 1 & 2)

		Study 1			Study 2		
	Mediator(s) - > y	Indirect Effect– Observer	Indirect Effect– Actor	Index of Moderated Mediation	Indirect Effect– Observer	Indirect Effect– Actor	Index of Moderated Mediation
Moderated Mediation Models [a]	1. NI - > Diag [b]	Significant	NS [c]	Significant	Significant	NS	Significant
	2. Diag - > OM [b]	Significant	Significant	Significant	Significant	NS	Significant
	3. Diag - > Trust	Significant	Significant	Significant	Significant	NS	Significant
	4. Diag - > Trusting Behaviors	Significant	Significant	Significant	Significant	NS	Significant
	5. NI - > Diag - > OM	Significant	NS	Significant	Significant	NS	Significant
	6. NI - > Diag - > OM - > Trust	Significant	NS	Significant	Significant	NS	Significant
	7. NI - > Diag - > OM - > Trust - > Trusting Behaviors	Significant	NS	NS	Significant	NS	Significant

[a] x = awareness of subsequent bad deed (yes vs. no), moderator = perspective (observer vs. actor).
[b] NI = Nefarious Intent, Diag = Diagnosticity, OM = Overall Morality.
[c] NS = Not Significant.

Appendix B. Alternative serial (Moderated) mediation models tested

		Study 1	Study 2	Study 3 [c]	Study 4	Study 5
		Index of Moderated Mediation	Index of Moderated Mediation	Indirect Effect	Indirect Effect	Indirect Effect
Alternative Serial (Moderated) Mediation Models	1. Diag - > NI - > OM [a]	NS [b]	Significant	N/A [d]	NS	Significant
	2. Diag - > NI - > OM - > Trust	NS	Significant	N/A [d]	NS	Significant
	3. Diag - > NI - > OM - > Trust - > Trusting Behaviors	NS	Significant	N/A [d]	NS	Significant
	4. OM - > NI - > Diag - > Trust	NS	NS	NS	NS	NS
	5. OM - > NI - > Diag - > Trust - > Trusting Behaviors	NS	NS	NS	NS	NS
	6. OM - > Diag - > NI - > Trust	NS	NS	NS	NS	NS
	7. Trust - > NI - > Diag - > OM	Significant	NS	Significant	NS	Significant
	8. OM - > Trust - > NI - > Diag	NS	NS	NS	NS	NS
	9. Trust - > OM - > NI - > Diag	NS	NS	Significant	NS	Significant

[a] Diag = Diagnosticity, NI = Nefarious Intent, OM = Overall Morality.
[b] NS = Not Significant.
[c] Because Study 3 manipulated nefarious intent, all of the alternative models were run without nefarious intent.
[d] Without nefarious intent, the model is the same as our predicted model.

References

Anderson, N. H. (1962). Application of an additive model to impression formation. *Science, 138*(3542), 817–818.
Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323–370.
Bell, B. E., & Loftus, E. F. (1989). Trivial persuasion in the courtroom: The power of (a few) minor details. *Journal of Personality and Social Psychology, 56*(5), 669–679.
Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin, 41*(4), 540–558.
Bottom, W. P., Gibson, K., Daniels, S., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science, 13*(5), 497–513.
Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372.
Cohen, T. R., & Morse, L. (2014). Moral character: What it is and what it does. *Research in Organizational Behavior, 34*, 43–61.
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 moral behavior. *Personality and Social Psychology Bulletin, 38*(7), 907–919.
Dirks, K. T., Kim, P. H., Ferrin, D. L., & Cooper, C. D. (2011). Understanding the effects of substantive responses on trust following a transgression. *Organizational Behavior and Human Decision Processes, 114*, 87–103.
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.
Ferrin, D. L., Kim, P. H., Cooper, C. D., & Dirks, K. T. (2007). Silence speaks volumes: The effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations. *Journal of Applied Psychology, 92*(4), 893–908.
Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Galeotti, F., Saucet, C., & Villeval, M. C. (2020). Unethical amnesia responds more to instrumental than hedonic motives. *Proceedings of the National Academy of Sciences, 117*(41), 25423–25428.
Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21–38.
Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394.
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.
Greenbaum, R. L., Mawritz, M. B., & Piccolo, R. F. (2015). When leaders fail to "walk the talk": Supervisor undermining and perceptions of leader hypocrisy. *Journal of Management, 41*(3), 929–956.
Haidt, J., & Kesebir, S. (2010). Morality. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social psychology* (5th ed., pp. 797–832). Wiley.
Harmon, D. H., Kim, P. H., & Mayer, K. J. (2015). Breaking the letter versus spirit of the law: How the interpretation of contract violations affects trust and the management of relationships. *Strategic Management Journal, 36*, 497–517.
Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). The Guilford Press.
Helzer, E. G., Furr, R. M., Hawkins, A., Barranti, M., Blackie, L. E. R., & Fleeson, W. (2014). Agreement on the perception of moral character. *Personality and Social Psychology Bulletin, 40*(12), 1698–1710.
Ito, T. A., Larsen, J. T., Smith, N. K., & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology, 75*(4), 887–900.
Janney, J. J., & Gove, S. (2010a). Reputation and corporate social responsibility aberrations, trends, and hypocrisy: Reactions to firm choices in the stock option backdating scandal. *Journal of Management Studies, 48*(7), 1562–1585.
Janney, J. J., & Gove, S. (2010b). Reputation and corporate social responsibility aberrations, trends, and hypocrisy: Reactions to firm choices in the stock option backdating scandal. *Journal of Management Studies, 48*(7), 1562–1585.
Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the moral landscape: Moral motives and group-based moralities. *Personality and Social Psychology Review, 17*(3), 219–236.
Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*(3), 521–537.
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.
Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107–128.
Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. Groups. *Organizational Behavior and Human Decision Processes, 120*(1), 1–14.
Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review, 34*(3), 401–422.
Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes, 99*, 49–65.
Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology vs. denial for repairing ability- vs. integrity-based trust violations. *Journal of Applied Psychology, 89*(1), 104–118.
Kim, P. H., & Harmon, D. H. (2014). Justifying one's transgressions: How rationalizations based on equity, equality, and need affect trust after its violation. *Journal of Experimental Psychology: Applied, 20*(4), 365–379.
Kim, P. H., Mislin, A., Tuncel, E., Fehr, R., Cheshin, A., & Van Kleef, G. A. (2017). Power as an emotional liability: Implications for perceived authenticity and trust after a transgression. *Journal of Experimental Psychology: General, 146*(10), 1379–1401.
Kim, P. H., Ployhart, R. E., & Gibson, C. B. (2018). Editors' comments: Is organizational behavior overtheorized? *Academy of Management Review, 43*(4), 1–5.
Kim, P. H., Wiltermuth, S. S., & Newman, D. (2021). A theory of ethical accounting and its implications for hypocrisy in organizations. *Academy of Management Review, 46*(1), 172–191.
Kouchaki, M., & Gino, F. (2015). Dirty deeds unwanted: The use of biased memory processes in the context of ethics. *Current Opinion in Psychology, 6*, 82–86.
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.
Laurent, S. M., Clark, B. A., Walker, S., & Wiseman, K. D. (2014). Punishing hypocrisy: The roles of hypocrisy and moral emotions in deciding culpability and punishment of criminal and civil moral transgressors. *Cognition & Emotion, 28*(1), 59–83. https://doi.org/10.1080/02699931.2013.801339
Leroy, H., Palanski, M. E., & Simons, T. L. (2012). Authentic leadership and behavioral integrity as drivers of follower commitment and performance. *Journal of Business Ethics, 107*(3), 255–264.
Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585–589.
Lyon, T. P., & Maxwell, J. W. (2011). Greenwash: Corporate environmental disclosure under threat of audit. *Journal of Economics and Management Strategy, 20*, 3–41.
Lyon, T. P., & Montgomery, A. W. (2013). Tweetjacked: The impact of social media on corporate greenwash. *Journal of Business Ethics, 118*(4), 747–757. https://doi.org/10.1007/s10551-013-1958-x
Maddux, W. W., Kim, P. H., Okumura, T., & Brett, J. M. (2011). Cultural differences in the function and meaning of apologies. *International Negotiation, 16*, 405–425.
Marín, L., Cuestas, P. J., & Román, S. (2016). Determinants of consumer attributions of corporate social responsibility. *Journal of Business Ethics, 138*(2), 247–260. https://doi.org/10.1007/s10551-015-2578-4
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.
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*(4), 494–498.
Montoya, A. K. (2018). *Conditional process analysis in two-instance repeated-measures designs [unpublished doctoral dissertation]*. Columbus, OH: The Ohio State University.
Montoya, A. K. (2019). Moderation analysis in two-instance repeated-measures designs: Probing methods and multiple moderator models. *Behavior Research Methods, 51*, 61–82.
Montoya, A. K., & Hayes, A. F. (2017). Two condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods, 22*(1), 6–27.
Morris, M. W., Menon, T., & Ames, D. (2001). Culturally conferred conceptions of agency: A key to social perception of persons, groups, and other actors. *Personality and Social Psychology Review, 5*(2), 169–182.
Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology, 67*(1), 363–385.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220.
Nisan, M. (1990). Moral balance: A model of how people arrive at moral decisions. In T. E. Wren (Ed.), *The moral domain: Essays in the ongoing discussion between philosophy and the social sciences* (pp. 283–314). The MIT Press.
Palanski, M. E., & Yammarino, F. J. (2007). Integrity and leadership: Clearing the conceptual confusion. *European Management Journal, 25*(3), 171–184.
Palanski, M. E., & Yammarino, F. J. (2011). Impact of behavioral integrity on follower job performance: A three-study examination. *The Leadership Quarterly, 22*(4), 765–786. https://doi.org/10.1016/j.leaqua.2011.05.014
Pizarro, D. A., & Tannenbaum, D. (2012). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer, & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). American Psychological Association.
Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86*(1), 61–79.
Rockwood, N. J. (2017). *Advancing the formulation and testing of multilevel mediation and moderated mediation models [Unpublished master's thesis]*. Columbus, OH: The Ohio State University.
Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology, 74*, 24–37.
Simons, T. L. (2002). Behavioral integrity: The perceived alignment between managers' words and deeds as a research focus. *Organization Science, 13*(1), 18–35. https://doi.org/10.1287/orsc.13.1.18.543
Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52*(4), 689–699.
Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131–142.
Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology, 47*(6), 1249–1254.
Tausch, N., Kenworthy, J. B., & Hewstone, M. (2007). The confirmability anbd disconfirmability of trait concepts revisited: Does content matter? *Journal of Personality and Social Psychology, 92*(3), 542–556.
Treviño, L. K., Hartman, L. P., & Brown, M. (2000). Moral person and moral manager: How executives develop a reputation for ethical leadership. *California Management Review, 42*(4), 128–142. https://doi.org/10.2307/41166057
Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.
Valente, M. J., & MacKinnon, D. P. (2017). Comparing models of change to estimate the mediated effect in the pretest-posttest control group design. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 428–450.
Wakslak, C. J., Jost, J. T., Tyler, T. R., & Chen, E. S. (2007). Moral outrage mediates the dampening effect of system justification on support for redistributive social policies. *Psychological Science, 18*(3), 267–274.
Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Sage Publications.