Cognition 236 (2023) 105437

Contents lists available at ScienceDirect

Cognition

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

Is it good to feel bad about littering? Conflict between moral beliefs and behaviors for everyday transgressions

Stephanie A. Schwartz, Yoel Inbar *

University of Toronto, Canada

ARTICLE INFO

Keywords:
Morality
Moral judgment
Metadesires
Character

ABSTRACT

People sometimes do things that they think are morally wrong. We investigate how actors' perceptions of the morality of their own behaviors affects observer evaluations. In Study 1 (n = 302), we presented participants with six different descriptions of actors who routinely engaged in a morally questionable behavior and varied whether the actors thought the behavior was morally wrong. Actors who believed their behavior was wrong were seen as having better moral character, but their behavior was rated as more wrong. In Study 2 (n = 391) we investigated whether perceptions of actor metadesires were responsible for the effects of actor beliefs on character judgments. We used the same stimuli and measures as in Study 1 but added a measure of the actor's perceived desires to engage in the behaviors. As predicted, the effect of actors' moral beliefs on judgments of their moral character was mediated by perceived metadesires. In Study 3 (n = 1092) we replicated these findings in a between-participants design and further found that the effect of actor beliefs on act and character judgments was moderated by participant beliefs about the general acceptability of the behavior.

A colleague[1] recently confessed to us that he routinely eats meat, even though he "knows" it is morally wrong to do so. Research suggests that he is not alone in this regard; many meat-eaters are "conflicted omnivores" who eat meat despite their moral reservations (Gendelman, 2017). Nor is conflict between moral ideals and behaviors limited to the food domain. Illegally downloading a movie, buying an article of clothing to wear it once and then return it, lying to avoid social awkwardness, driving alone rather than taking public transit—many of us have done one or more of these things despite believing it is at least a little wrong to so. In fact, often most unethical behavior is done by people who only transgress a little, and so can see themselves as ethical despite some minor unethicality (Mazar, Amir, & Ariely, 2008).

How do observers make sense of someone who acts contrary to their moral standards? Because these cases involve a conflict between moral ideals and actions, they are likely to be particularly influential in judgments of moral character (i.e., holistic judgments of a person's inner moral essence; Hartman, Blakey, & Gray, 2022). Recent research in person perception and moral psychology has highlighted the importance of these judgments. Warmth and competence had long been thought to be the primary dimensions underlying our holistic judgments of others (Fiske, Cuddy, & Glick, 2007; Fiske, Cuddy, Glick, & Xu, 2002), but perceivers' impressions of a target's morally relevant traits (e.g., fairness, honesty, courage, or loyalty) predict overall person judgments more strongly than do perceptions of traits related to either warmth (Goodwin, Piazza, & Rozin, 2017) or competence (Wojciszke, Bazinska, & Jaworski, 1998). Moral qualities thus seem to be what people weight most heavily in their impressions of others, likely because they are seen as especially predictive of future interpersonal behavior (Hartman et al., 2022; Uhlmann, Pizarro, & Diermeier, 2015).

Likewise, moral psychology has recently focused more on the importance of moral character judgments. Influential early theories of moral judgment investigated moral blame for acts, focusing on perceptions of harm, controllability, and intentionality (Darley & Zanna, 1982; Shultz & Schleifer, 1983; Shultz, Schleifer, & Altman, 1981; Shaver & Drown, 1986; Weiner, 1995). However, subsequent research has shown that moral judgments are often responsive to the perceived character of the actor rather than to the act in isolation. When people encounter information about moral or immoral behaviors, they also infer what those behaviors signal about the actor's character (Uhlmann et al., 2015). This can lead to "act-person" dissociations in which committing an act that is acceptable or even required by a normative rule might also reveal something negative about the actor's character. For example, in

* Corresponding author at: Department of Psychology, University of Toronto, 1265 Military Trail, Toronto, Ontario M1C 1A4, Canada.
E-mail address: yoel.inbar@utoronto.ca (Y. Inbar).

[1] Stéphane Côté, who (perhaps ironically) is the Geoffrey Conway Chair in Business Ethics at the Rotman School of Management.

https://doi.org/10.1016/j.cognition.2023.105437
Received 30 June 2022; Received in revised form 4 March 2023; Accepted 6 March 2023
Available online 28 March 2023

many circumstances people think consequentialist decisions (e.g. choosing to sacrifice one person to save five) are morally correct (because they follow the moral rule "save the most people possible"). However, people also often see the person who actually chooses to sacrifice one to save five as *less* moral (because sacrificing someone shows willingness to inflict harm and a lack of empathy; Uhlmann, Zhu, & Tannenbaum, 2013). Act-person dissociations can also arise for moral transgressions. For example, people think that it is morally worse to beat up one's girlfriend than to beat up her cat, but also see a cat-beater as morally worse than a girlfriend-beater, possibly because cat-beating demonstrates a particularly extreme lack of empathy (Tannenbaum, Uhlmann, & Diermeier, 2011).

The fact that act and character judgments seem to follow different rules naturally raises the question as to what those rules are. Researchers have proposed that perceivers make moral character judgments by inferring how much an actor possesses socially-essential dispositions such as empathy, compassion, and trustworthiness (Helzer & Critcher, 2018). These traits are particularly socially valued because they signal that the actor will be a reliable partner in future interactions (Anderson, Crockett, & Pizarro, 2020; Critcher, Helzer, & Tannenbaum, 2020; Helzer & Critcher, 2018; Uhlmann et al., 2015). To make these judgments, observers evaluate not only people's actions, but also what seems to have motivated them (i.e., whether people had the "right" moral motivations; Critcher et al., 2020; Uhlmann et al., 2013). When underlying motivations seem to conflict with actions, people adjust their blame and praise judgments in line with the inferred motivation. For example, someone who refrains from immoral behavior but is tempted by it is judged more negatively than someone who is not tempted (Berman & Small, 2018). Likewise, someone who engages in praiseworthy behavior only after deliberation is praised less than someone who acts without hesitation, because observers infer mixed motives in the former case (Critcher, Inbar, & Pizarro, 2013).

A special kind of motive that is particularly important for moral judgments is a "metadesire," which is defined as a preference for one of two conflicting desires or impulses over another (Frankfurt, 1973, 1987). For example, an addict might have a desire to use drugs (because of his physical cravings) and a conflicting desire to stop using. If he prefers that the desire to stop using drugs win out, he has a metadesire not to be an addict. People seem to treat metadesires as an indication of what actors "really" want and adjust their moral judgments accordingly. For example, wrongdoers who are inferred to have metadesires inconsistent with their immoral actions (for example, someone who becomes violent in a fit of rage) are judged less harshly (Pizarro, Uhlmann, & Salovey, 2003).

Perceived metadesires may be particularly important in judgments of people who act contrary to their moral standards. Consider again the conflicted omnivore, who eats meat despite thinking it wrong to do so. It seems plausible to infer that such a person would like to see his desire to avoid eating meat win out over his desire to indulge. And, in general, it may be that people believe that agents who act in conflict with their moral standards do not truly want to do what they are doing. Thus, doing something one believes to be wrong may signal something positive about one's moral character—that one has moral standards (even if one is currently not living up to them). We therefore predict that holding behavior constant, actors who thinks their behavior is morally wrong will be judged to have better moral character than actors who do not.

What about judgments of the *act* itself, rather than the actor? Here, there are reasons to think that an act will be condemned *more* when the person engaging in it thinks it immoral. This could be for two reasons. First, an actor's belief about the behavior might be informative to observers about its moral status. Considering the relatively benign nature of behaviors such as meat-eating, it is likely that evaluations of the acceptability of acts will be neither extremely positive nor negative, and therefore leave room for malleability in perceptions. If this is the case, an actor believing the act to be morally wrong may shift people's perceptions of the act itself. If this is the case, learning that the actor considers their behavior to be immoral should change perceptions of the moral acceptability of the behavior in general.

Second, people may (implicitly or explicitly) hold the view that people should not do things that *they themselves* believe to be morally wrong (regardless of whether, normatively, they are wrong or not). If this is the case, learning that the actor considers the behavior to be immoral should shift perceptions of the moral acceptability of the actor's behavior, but not necessarily of the act in general. To distinguish these possibilities, we asked participants to rate both the immorality of the actor's behavior, and the moral acceptability of the act in general.

1. The current research

Previous research on character judgments has often asked people to evaluate extreme or unusual behaviors such as pushing an injured passenger off a lifeboat (Uhlmann et al., 2013), calling in a missile strike on a terrorist meeting (Critcher et al., 2020), or beating a girlfriend's pet cat (Tannenbaum et al., 2011). These unusual or unrealistic scenarios are analogous to the geneticist's fruit flies—they are simplified cases that nonetheless illuminate the mental processes underlying moral judgments (Greene, 2009). At the same time, using these sorts of unrealistic cases to study morality assumes that they evoke judgment processes that are the same as those used to evaluate the kinds of behaviors people encounter in everyday life. Some researchers have argued strongly this is not the case, and that this has led moral psychologists to draw incorrect conclusions (Gray & Keeney, 2015). This problem may be particularly acute for judgments of character, which (it is argued) are motivated by the everyday social problem of deciding who is a reliable interaction partner worthy of trust and investment (Helzer & Critcher, 2018). It may be that the processes that are used to make sense of the moral character of military commanders or shipwreck survivors are not the same as those used in everyday social judgment—which is ultimately what researchers are trying to explain. In the current research, we therefore examine everyday behaviors, that is, mild moral transgressions common enough that many people engage in them or observe them at least sometimes.

In three studies, we investigated how actors' perceptions of the morality of their own behaviors affect evaluations of their character and of their actions. In Study 1, we presented participants with six different descriptions of actors who routinely engaged in a morally questionable behavior and varied whether the actors thought the behavior was morally wrong. We predicted that actors who believed their behavior was wrong would be seen as having better moral character, but that the act itself would be rated as *more* wrong.

In Study 2 we tested whether perceptions of actor metadesires are responsible for the effects of actor beliefs on character judgments. We used the same stimuli and measures as in Study 1 but added a measure of the actor's perceived metadesires regarding the behaviors. We predicted that perceptions that of actor metadesires would mediate the relationship between the actors' beliefs regarding the morality of their actions and the perceptions of their moral character.

Studies 1 and 2 used within-participant designs in which participants read about actors who thought their behavior was morally wrong, actors who did not, and control actors for whom no belief information was provided. This design maximizes statistical power but it may also make actor beliefs especially salient to participants. In Study 3, we therefore turned to a between-participants design in which each participant read only one scenario describing a single actor. We also examined perceived extremity of the action as a possible boundary condition by testing whether participants' own beliefs about the general acceptability of the behavior moderated effects of actor beliefs on judgments.

2. Study 1

In Study 1 we tested the effect of individuals' private beliefs about the acceptability of their own actions on judgments of their actions and moral character.

2.1. Immoral behavior generation and pre-test

We generated candidate immoral behaviors ourselves and with an open-ended survey in which we asked 20 Amazon Mechanical Turk workers to list any immoral behaviors they had witnessed or committed in the last 24 h. From this list, we selected 20 behaviors and verified that none were seen as extremely immoral or uncommon by asking a second group of Mechanical Turk workers ($n = 99$) to rate either the wrongness of each behavior ("To what extent do you believe that [behavior] is immoral?") or the frequency with which they personally engaged in it ("With what frequency do you personally engage in [behavior]?") on 100-point slider scales. The highest mean immorality rating was 60.73 (for "texting which driving") and the lowest frequency rating was 13.94 ("running a stop sign or red light while driving"). We removed four behaviors that, after discussion, we thought might perform badly in the main studies, leaving us with a final list of 16 (see Table 1).[2] Full descriptions of each behavior and of the pilot studies are available in the Supplemental Materials.

3. Method

3.1. Participants

401 United States residents (203 Female, 190 Male, 8 reporting "another" gender, $M_{age} = 31.69$, $SD_{age} = 11.53$) were recruited from the web recruitment platform Prolific.co in October of 2020. Participants completed the study via a link to a Qualtrics survey in the Prolific interface. Participants were compensated $0.53 for their time.

3.2. Materials and procedure

We created three short scenarios for each of the 16 behaviors that always described an actor performing the behavior but varied the actor's beliefs. In the control versions, no information was given about the actor's beliefs about the behavior. In the *not wrong* versions, actors were described as believing that it was not wrong to engage in the behavior. In the *wrong* versions, actors were described as engaging in the behavior despite believing that it is wrong to do so. For example, for the behavior "speeding while driving," the three versions were:

Control: "Sam often speeds while driving."

Not wrong: "Sam often speeds while driving. He does not believe that it is wrong to do so."

Wrong: "Sam often speeds while driving, although he believes that it is wrong to do so."

This resulted in 48 scenarios total (16 × 3). Participants completed a survey hosted on Qualtrics that presented them with six behaviors: two control, two *not wrong*, and two *wrong*. Each participant was presented with a set of six behaviors, with two behaviors in each of the three conditions.[3] After reading each scenario, participants were asked to answer five questions. Two focused on judgments of the act: "How wrong is it for [protagonist] to [act]?" and "How immoral is it for [protagonist] to [act]?" both (1 = *Not at all*, to 7 = *Completely*). Two focused on the actor's character: "Based on what you read, do you think [protagonist] is mainly a good person or a bad person?" (1 = *Mainly a bad person*, to 7 = *Mainly a good person*), and "Based on what you read, do you think [protagonist] has good moral standards?" (1 = *Not at all*, to 7 = *Completely*). A final question assessed participants' beliefs about the moral acceptability of each behavior in general (*general moral acceptability*): "How morally acceptable do you, personally, think it is to [act]?" (1 = *Completely unacceptable,* to 7 = *Completely acceptable*). We included this question to be able to distinguish participants' agent-specific act judgments from their beliefs about the moral acceptability of the behavior in general.

The last section of the survey consisted of demographic questions and two attention checks. The first attention check was formatted identically to the behavioral descriptions in the previous section. Participants were instructed, "This is an attention check. Please choose 'Not at all' for each of the two questions below." The second attention check, which was intended to gauge attentiveness and English fluency, read "Please briefly describe, in your own words, what you were asked to do in this study," with a text entry field for open-ended responses.

Table 1
Mean pilot ratings of immorality and self-reported frequency of 16 behaviors used in Studies 1 and 2 (between participants).

	Frequency rating (0–100)		Immorality rating (0–100)	
Behavior	Mean	SD	Mean	SD
Eating meat	65.28	29.94	14.69	22.59
Speeding (while driving a car)	34.24	26.31	47.33	33.03
Texting while driving a car	22.06	27.08	60.73	36.03
Gossiping	25.12	23.6	38.35	29.31
Littering	14.94	21.78	56.67	33.38
Not recycling recyclable waste	41.26	30.78	38.94	30.79
Illegally downloading TV shows, music, or movies	26.32	32.42	38.92	29.94
Swearing / cursing	49.02	29.9	21.33	29.24
Telling "white" / small lies	28.78	24.6	36.47	27.67
Avoiding giving money to the homeless	42.02	29.52	30.57	31.27
Slacking off / not working your hardest while at work	30.36	25.93	40.18	32.32
Running a stop sign or red light while driving	13.94	20.72	57.57	31.63
Drinking bottled water	51.44	33.75	18.16	27.61
Leaving dirty dishes or garbage in communal areas at home / school / work	21.58	24.23	43.53	31.77
Spending money frivolously	28.32	24.88	33.80	32.82
"Checking out" strangers	36.20	27.33	25.55	30.33

4. Results

Responses to the open-ended attention check questions were coded by two independent research assistants. Each response was coded for English fluency and for correctness of the response in describing the nature of the study. A third evaluator was used to break the tie for any disagreements between the first two raters. We removed 99 participants for lack of English fluency, inaccurate descriptions of what they had just done, or failing the closed-ended attention check, leaving a total of 302 participants in the final data set. The two act-focused questions were highly correlated for each behavior ($rs = 0.66$–0.89) as were the two character-focused questions ($rs = 0.68$–0.87). Additionally, exploratory factor analyses showed a two-factor solution with both act questions loading on one factor and both character questions loading on another for every behavior. We therefore created an *act wrongness* composite by averaging the two act-focused questions, and a *moral character* composite by averaging the two character-focused questions.

Psychologists typically analyze dependent data (e.g., data where the same individual responds to multiple stimuli) using mixed-effects (i.e., multilevel) models. However, methodologists have recommended the use of simpler fixed-effects models with clustered standard errors if researchers are simply trying to account for non-independence in the data (McNeish, Stapleton, & Silverman, 2017). We therefore use OLS

[2] Three of these were specific to urban living (e.g., "Driving somewhere within walking distance") and we thought that urbanites might respond very differently to them than non-urbanites. The last was an infrequent inaction ("Not donating blood") which made the items describing it read oddly.

[3] Each participant completed a subset of six of the 48 possible condition X behavior combinations. Participants were randomly assigned to blocks of six question sets. Each block was constructed semi-randomly to include six unique behaviors, with two falling into each of the three conditions.

regression with fixed effects for scenario and standard errors clustered by participant, but results are very similar using multilevel modeling.

We separately modeled act wrongness and moral character composites as a function of condition with fixed effects for scenario and cluster-robust standard errors by participant.[4] We tested two versions of each of these models. The first included no covariates. The second covaried *general moral acceptability* ratings of the act to control for a participant's own beliefs about a given behavior's moral status. Results from these models are shown in Table 2, and mean act wrongness and moral character ratings per condition are shown in Figs. 1 and 2.

Compared to the control condition (where participants were given no information about actor beliefs), act wrongness ratings were significantly higher in the *wrong* condition. There was no difference in act wrongness ratings between the control and *not wrong* conditions. That is, when actors believed their actions to be wrong, they were seen as more wrong. However, when actors did not believe their actions to be wrong, they were not seen as less wrong relative to the control condition (see Fig. 1).

Moral character ratings showed a different pattern: When actors believed their actions to be wrong, their character was rated more *positively* (relative to the control condition). When they did not believe their actions to be wrong, their character was rated more negatively (again, relative to the control condition). In other words, perceptions of wrongdoers' character respond positively to their belief that they are doing something morally wrong (and negatively to their belief that they are not; see Fig. 2).

To investigate whether actor beliefs affect participants' ratings of the general moral acceptability of acts, we regressed act acceptability (e.g., "How morally acceptable do you, personally, think it is to litter?") onto condition. Results showed a small effect of condition on personal acceptability such that participants rated the behaviors as less personally acceptable compared to control when the actor in the vignettes also viewed the behaviors as wrong, $t(1{,}793) = -2.03$, $p = .04$, control vs. *not wrong*, $t(1{,}793) = -1.02$, $p > .1$, *not wrong* vs. *wrong*, $t(1{,}793) = -0.36$, $p > .1$. However, including personal acceptability ratings in these models did not affect the relationships between condition and actor's act or character ratings (see Table 2).

5. Study 2: Mediation by perceived metadesires

Study 2 had two primary goals. First, we wanted to replicate the divergent effects of actor beliefs about wrongness on act and character evaluations that we found in Study 1. Second, we wanted to test whether the effects of actor wrongness beliefs on evaluations was mediated by perceived "metadesires." Metadesires, also called "second-order desires" (Frankfurt, 1973, 1987), are defined as an agent's acceptance or rejection of first-order desires. Previous research has shown that perceived metadesires affect both blame for negative acts and praise for positive ones (Pizarro et al., 2003). In the current case, we hypothesized that actor beliefs about act wrongness would affect perceived metadesires. Compared to the control condition (where participants did not see any information about the actor's beliefs) we predicted that actors who saw their act as wrong would be seen as having metadesires inconsistent with the act. Conversely, we predicted that actors who did not see their act as wrong would be seen as having metadesires consistent with the act. We further predicted that differences in perceived metadesires would mediate the relationship between actor beliefs about wrongness and evaluations of their character. Study 2 was therefore a direct replication of Study 1 with the addition of measures of perceived metadesires of the actors presented in our vignettes. Unless noted otherwise, sample size, measures, and analyses were preregistered at AsPredicted.org (https://aspredicted.org/fq8fh.pdf).

6. Method

6.1. Participants

We recruited 403 United States residents from Prolific.co in January of 2022. After excluding 10 participants who failed our preregistered attention check as well as 39 participants with incomplete data, we were left with 391 (258 Female, 124 Male, 9 reporting "another" gender, and 2 failing to report gender, $M_{age} = 33.28$, $SD_{age} = 11.71$). Participants completed the study via a link to a Qualtrics survey in the Prolific interface. Participants were compensated $1.01 for their time.

6.2. Materials and procedure

Study 2 used the same survey as Study 1, with two modifications. Again, participants were asked to read six scenarios out of a possible 48 (16×3): two control, two *not wrong*, and two *wrong*. For example, for the behavior "littering," the three versions were:

Control: "Kyle occasionally litters."

Not wrong: "Kyle occasionally litters. He does not believe that it is wrong to do so."

Wrong: "Kyle occasionally litters, although he believes that it is wrong to do so."

The first modification was the addition of three questions to assess perceived actor metadesires. Participants were asked the extent to which they agreed with each of the following: "[Protagonist] wishes s/he **didn't** [act]," "[Protagonist] really wants to [act]," and, "[Protagonist] feels good about [act]." For example, in the "littering" scenario, these questions read "Kyle wishes he **didn't** litter," "Kyle really wants to litter," and "Kyle feels good about littering."[5]

The second modification made to Study 2 was the removal of the open-ended attention check. This was done because the removal of participants who failed the attention check did not alter our findings in Study 1 and the coding of the open-ended attention check required a large amount of research assistant time. The attention check items used in Study 2 were: "Sandy loves to pay attention to studies she participates in. Please choose 'Not at all' for each of the two questions below," followed by two questions that appeared similar to the behavior ratings used throughout the survey.

7. Results

We again created an *act wrongness* composite by averaging the two act-focused questions ($\alpha = 0.92$) and a *moral character* composite by averaging the two character-focused questions ($\alpha = 0.90$). Additionally, we created a *metadesires* composite by first reverse coding the item, "[Protagonist] wishes s/he **didn't** [act]," and then averaging this item with the other two metadesire questions ($\alpha = 0.83$). Responses scales were numbered such that higher numbers indicates *lower* perceived endorsement of the act by the actor, i.e., more positive perceived metadesires.

As in Study 1, we again separately modeled act wrongness and moral

[4] These were implemented using the *sandwich* package in R (Zeileis, 2004; Zeileis, Köll, & Graham, 2020). We specified bias adjustment HC1 (the default), which is a degrees of freedom-based correction.

[5] Metadesires (or, as he called them, "second-order volitions") were originally defined by Frankfurt (1973) as a preference between two conflicting first-order desires. Having a metadesire means that a person "identifies" themselves with one desire over the other and wants it to be "effective" (i.e., action-guiding). This means that someone who acts inconsistently with their metadesires does not truly want to do what they are doing. Because in Frankfurt's conception desires are seen as closely linked to action, and because we thought it would be clearer to participants to ask about identification with acts than desires, we phrased our questions to ask about identification with the act directly (rather than the desire leading to the act).

Table 2
Effects of condition on judgments of act wrongness and actors' moral character (Study 1). Both models include fixed effects for behavior and cluster-robust standard errors (clustered by participant). Model specification 1 includes no further control variables. Model specification 2 controls for participants' beliefs about the general moral acceptability of each behavior.

	Model Specification 1					
	Act wrongness			Moral character		
	Unstandardized *B* [95% CI]	*t*(1,793)	*p*	Unstandardized *B* [95% CI]	*t*(1,793)	*p*
Condition						
Baseline – Control	–	–	–	–	–	–
Not Wrong	0.002 [−0.10, 0.10]	0.04	0.97	−0.17 [−0.27, −0.06]	−3.11	0.002
Wrong	0.15 [0.05, 0.24]	3.02	0.003	0.11 [0.01, 0.21]	2.07	0.04

	Model Specification 2					
	Act Wrongness			Moral Character		
	Unstandardized *B* [95% CI]	*t*(1,793)	*p*	Unstandardized *B* [95% CI]	*t*(1,793)	*p*
Condition						
Baseline – Control	–	–	–	–	–	–
Not Wrong	−0.04 [−0.11, 0.04]	−0.10	0.32	−0.13 [−0.22, −0.05]	−3.04	0.002
Wrong	0.08 [0.01, 0.16]	2.17	0.03	0.16 [0.07, 0.26]	3.45	< 0.001

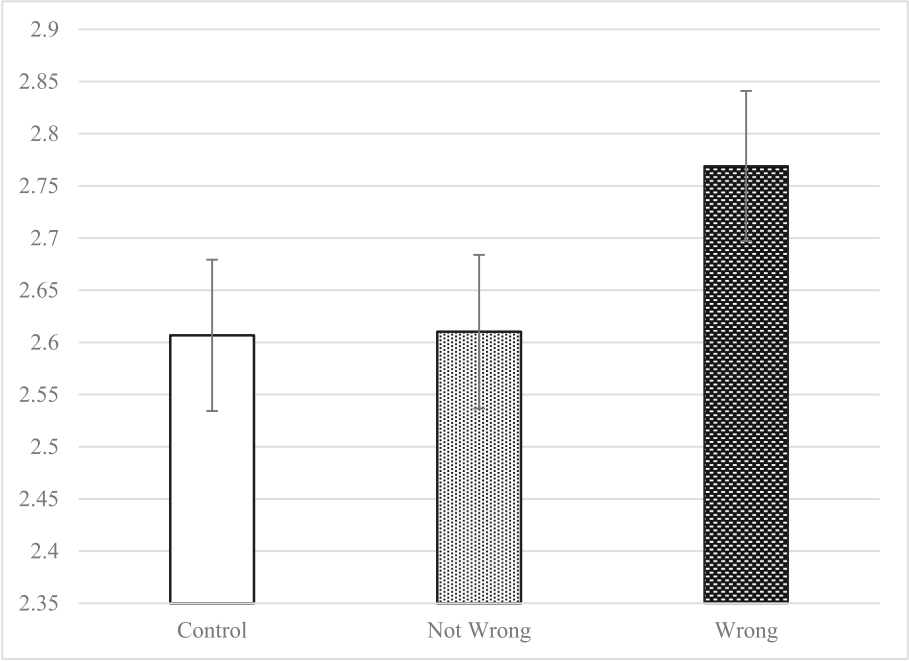

Fig. 1. Average ratings of act wrongness by condition in Study 1.
Note. Higher ratings indicate greater act wrongness. Error bars show standard errors.

character composites as a function of condition with fixed effects for scenario and cluster-robust standard errors by participant. As in Study 1, we tested two versions of each of these models. The first included no covariates. The second covaried *general moral acceptability* ratings of the act to control for a participant's own beliefs about a given behavior's moral status (note that this second version was not preregistered; it is included only as a robustness check). Results from these models are consistent across specifications (see Table 3). Mean act wrongness and moral character ratings per condition are shown in Figs. 3 and 4.

Compared to the control condition, act wrongness ratings were again significantly higher in the *wrong* condition. Again, there was no difference in act wrongness ratings between the control and *not wrong* conditions, replicating our findings from Study 1 that when actors believed their actions to be wrong, they were seen as more wrong. However, when actors did not believe their actions to be wrong, they were not seen as less wrong relative to the control condition (see Fig. 3).

Moral character ratings replicated the pattern found in Study 1. When actors believed their actions to be wrong, their character was rated more positively than in the control condition. When actors did not believe their actions to be wrong, their character was rated more negatively than in the control condition (see Fig. 4). Thus, for both act and character ratings, the results of Study 2 fully replicate Study 1. Additionally, perceived metadesires of actors also significantly varied by condition, such that when actors viewed their action as wrong, participants perceived their desire to not engage in that behavior as higher than both actors in the control and not wrong conditions. Further, actors in the *not wrong* condition were assumed to have even weaker intentions to *not* engage in the immoral actions than those in the control conditions (Fig. 5).

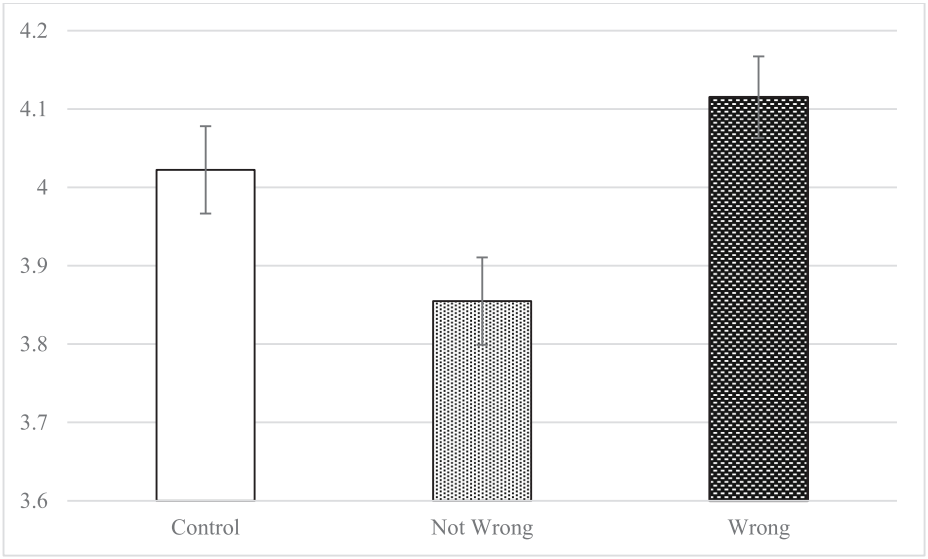

Fig. 2. Average ratings of actors' moral character by condition in Study 1.
Note. Higher ratings indicate more positive moral character. Error bars show standard errors.

Table 3
Effects of condition on judgments of act wrongness, actor moral character, and actor metadesires (Study 2). Both models include fixed effects for behavior and cluster-robust standard errors (clustered by participant). Model specification 1 includes no further control variables. Model specification 2 controls for participants' beliefs about the general moral acceptability of each behavior.

Model Specification 1

	Act Wrongness			Moral Character			Metadesires		
	Unstandardized *B* [95% CI]	*t* (2,301)	*p*	Unstandardized *B* [95% CI]	*t* (2,301)	*p*	Unstandardized *B* [95% CI]	*t* (2,301)	*p*
Condition									
Baseline – Control	–	–	–	–	–	–	–	–	–
Not Wrong	0.04 [−0.06, 0.13]	0.73	0.47	−0.22 [−0.32, −0.12]	−4.39	< 0.001	−0.57 [−0.65, −0.49]	−14.30	< 0.001
Wrong	0.15 [0.06, 0.24]	3.07	0.002	0.36 [0.26, 0.45]	7.46	< 0.001	1.07 [0.98, 1.15]	24.58	< 0.001

Model Specification 2

	Act Wrongness			Moral Character			Metadesires		
	Unstandardized *B* [95% CI]	*t* (2,301)	*p*	Unstandardized *B*	*t* (2,301)	*p*	Unstandardized *B* [95% CI]	*t* (2,301)	*p*
Condition									
Baseline – Control	–	–	–	–	–	–	–	–	–
Not Wrong	0.01 [−0.06, 0.08]	0.30	0.77	−0.20 [−0.27, −0.12]	−5.00	< 0.001	−0.57 [−0.65, −0.49]	−14.36	< 0.001
Wrong	0.12 [0.05, 0.19]	3.45	< 0.001	0.38 [0.30, 0.47]	9.04	< 0.001	1.07 [0.99, 1.16]	24.84	< 0.001

7.1. Mediation by perceived metadesires

We used the R package *mediation* (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014) to estimate the indirect effect of condition on character ratings via perceived metadesires. This package estimates indirect effects, 95% confidence intervals, and *p*-values using a quasi-Bayesian Monte Carlo method based on normal approximation (Imai, Keele, & Tingley, 2010). We specified the default number of simulations (1000). Because condition is a dummy-coded variable with three levels, we present separate models for the comparison between: (1) the control and the *not wrong* conditions; (2) the *wrong* and *not wrong* conditions.

7.1.1. Character ratings

The effect of actors' beliefs about the wrongness of their own behaviors on participants' ratings of the actor's character was fully mediated by the perceived metadesires of the actors in the vignettes (i.e., how they were perceived to *want* to behave). In other words, the relationship between (e.g.) Kyle's views on his own littering and participants' views on his moral character was fully explained by participants' assumptions regarding Kyle's actual desire to [not] litter. This effect was found both when comparing the control condition to the *not wrong* condition, as well as the *not wrong* to the *wrong* condition.

As Fig. 6 illustrates, using the control condition as the baseline, the regression coefficient between the actor's view that their actions were wrong, and perceptions of actor character, as well as the regression coefficient between perceived actor metadesires and perceptions of actor character were significant. The average unstandardized indirect effect was 0.33 [95% CI: 0.26, 0.41], $p < .001$. The remaining effect of

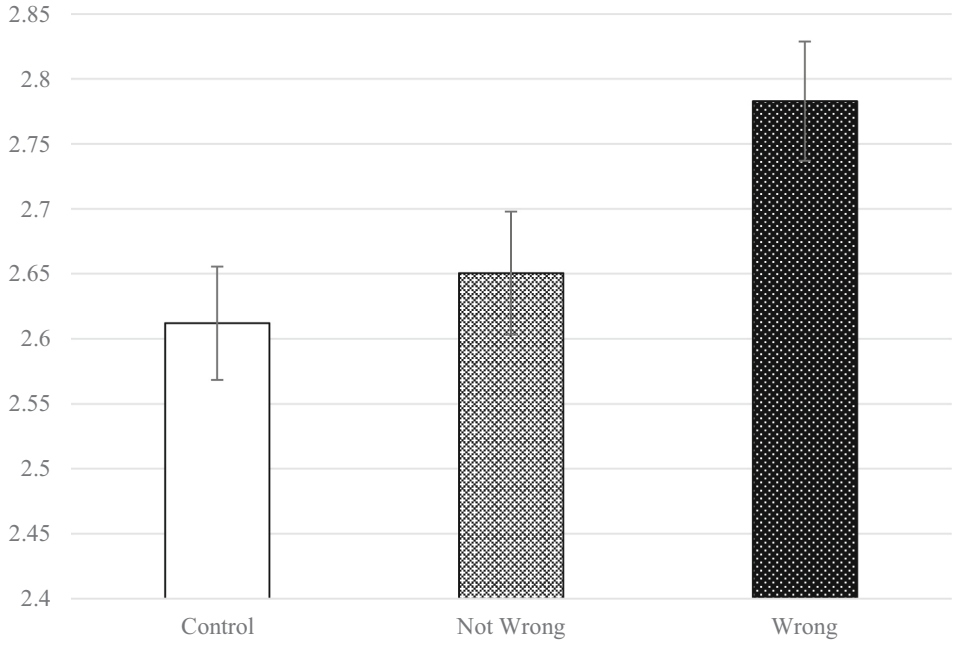

Fig. 3. Average ratings of act wrongness by condition in Study 2.
Note. Higher ratings indicate greater act wrongness. Error bars show standard errors.

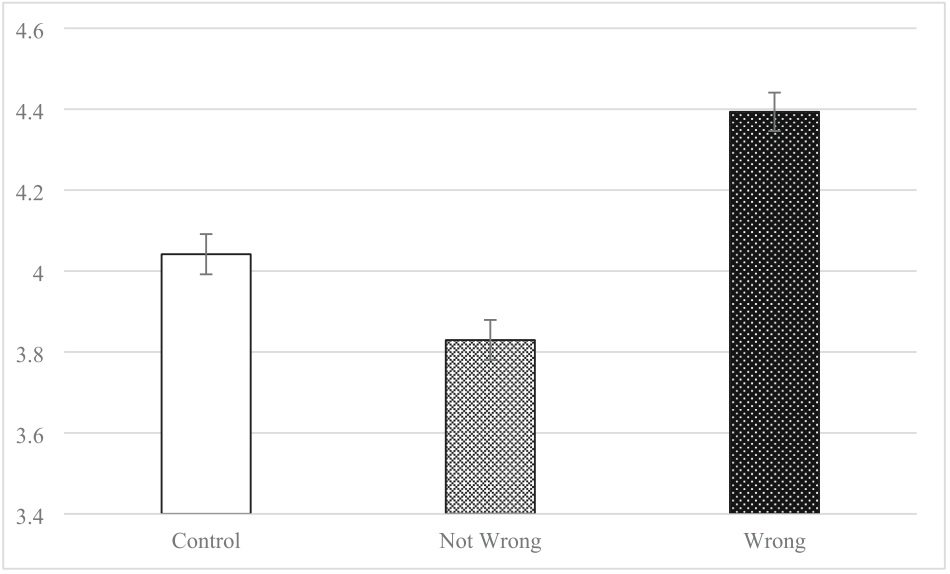

Fig. 4. Average ratings of actors' moral character by condition in Study 2.
Note. Higher ratings indicate more positive moral character. Error bars show standard errors.

the actor's view of their own behaviors on perceptions of their character not accounted for by perceived metadesires was no longer significant, (residual effect = 0.03, $p > .1$), indicating full mediation of the relationship between actor beliefs and their perceived character by their perceived metadesires.

Fig. 7 illustrates the same analysis but testing the effect of the *wrong* condition vs. the *not wrong* (baseline) condition. Here, the average indirect effect of condition via perceived metadesires was 0.51 [95% CI: 0.41, 0.63], $p < .001$. The remaining effect of the actor's view of their own behaviors on perceptions of their character not accounted for by perceived metadesires was no longer significant, (residual effect = 0.06, $p > .1$), indicating full mediation of the relationship between actor's belief and their perceived character by their perceived metadesires.

7.2. Overall act evaluations

Unlike in Study 1, we found no effect of an actor's beliefs about the acceptability of their actions on participants' judgments of the general moral acceptability of the act, all *ps* > 0.10.

8. Study 3

In Studies 1 and 2, participants saw two scenarios from each condition (*wrong*, *not wrong*, and control), meaning that they saw six in all. This design maximized power by allowing us to control for between-participant variability, but it may have made actor beliefs particularly salient to participants. In Study 3, we tested whether the effect of actor beliefs on judgments would emerge in a between-participants design in

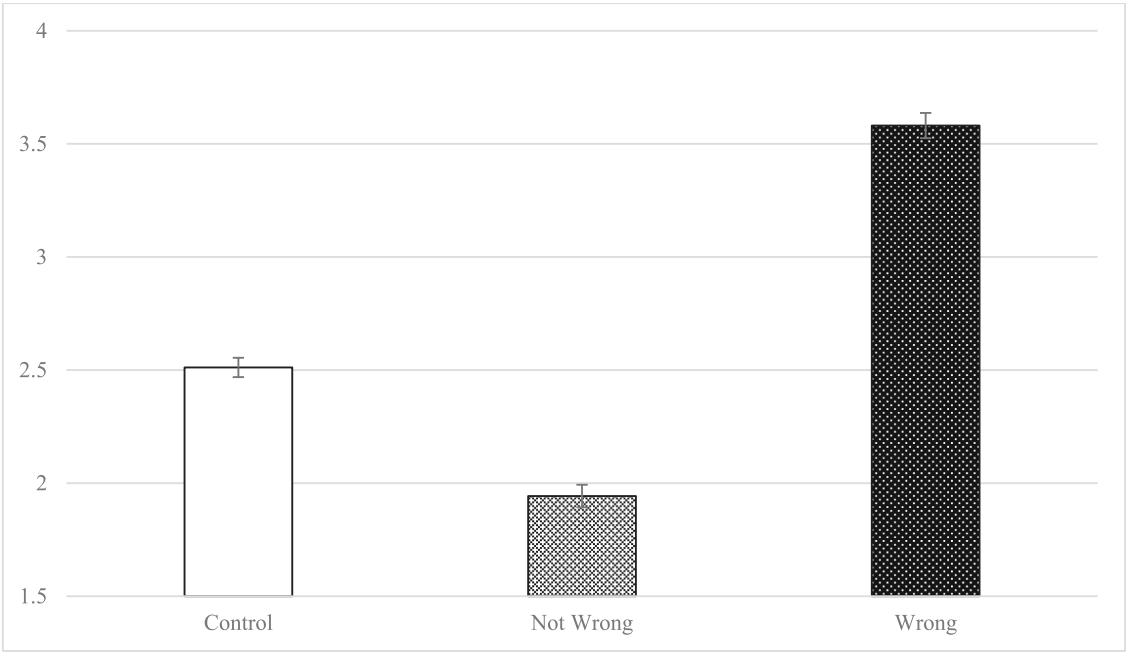

Fig. 5. Average metadesire ratings by condition in Study 2.
Note. Higher numbers indicate stronger beliefs that an actor wishes they did not engage in the described behaviors (i.e., more positive metadesires). Error bars show standard errors.

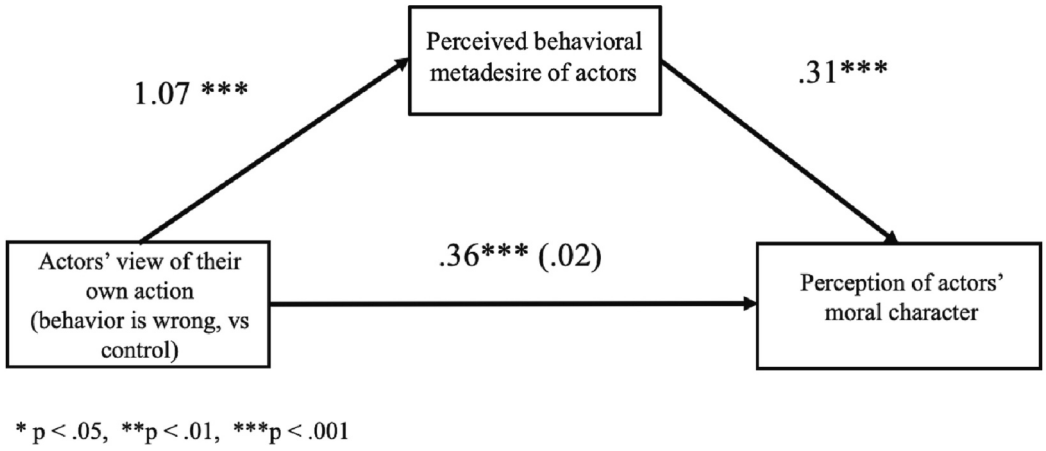

Fig. 6. Mediation of the relationship between actors' own view of their actions and perception of actors' moral character by perceived behavioral metadesires of actors, comparing the control condition to the not wrong condition.
Note. Brackets indicate direct effect controlling for mediation

which each participant read only one scenario.

For theoretical reasons, we have focused specifically on everyday behaviors rather than extreme or unusual moral transgressions. It may therefore be that the extremity of the transgression is a boundary condition on the effects we observed. One way to test this question would be to use a new, more extreme set of immoral acts. However, examining participant ratings of overall act acceptability showed that across participants and acts, there was substantial variability in how morally acceptable participants found the acts overall. For example, in Study 2 25% of moral acceptability ratings were 2 or below and 25% were 5 or above (recall that ratings were on a 1–7 scale anchored by *Completely unacceptable* and *Completely acceptable*). We therefore reasoned that we could test the moderating effects of transgression extremity using the current set of acts by asking participants to rate the general moral acceptability of the act *before* they saw information about a specific actor. This allowed us to test whether actor beliefs were more or less influential in judgments when participants saw the act as less morally acceptable in general. For character judgments, there are two plausible but opposite predictions. First, it may be that engaging in a highly immoral act is so damning that character judgments no longer respond to actor beliefs, which would predict a smaller effect of condition among those who find the act highly immoral. On the other hand, it may be that actor beliefs become even more important in character judgments for the most severe offenses—imagine, for example, your reaction to an unrepentant murderer compared to a repentant one (in fact, jurors report heavily weighing remorse or its absence when deciding whether to sentence convicted murderers to die; Haney, Sontag, & Costanzo, 1994). This would predict a *larger* effect of condition for observers who thought that the act in general was particularly immoral. Because we thought either prediction was plausible, we did not make a directional prediction for character judgments.

What about act judgments? In Studies 1 and 2, we found that acts

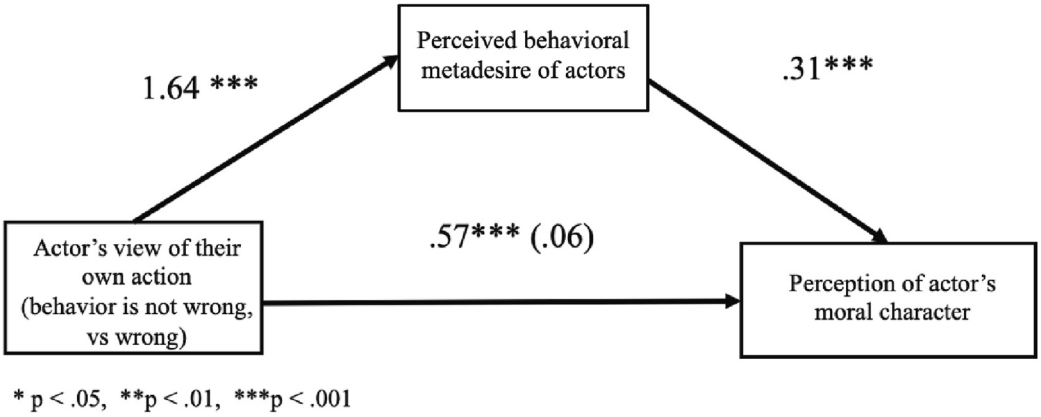

Fig. 7. Mediation of the relationship between actors' own view of their actions and perception of actors' moral character by perceived behavioral metadesires of actors, comparing the wrong condition to the not wrong condition.
Note. Brackets indicate direct effect controlling for mediation

were condemned more when the people engaging in them believed them to be immoral. Beliefs about the morality of the act *in general* were not consistently affected by actor beliefs about wrongness, suggesting that greater condemnation was motivated by observer beliefs that people should not do things that they themselves think are morally wrong (that is, they should not violate their own moral standards). We reasoned that if this is the case, then actor beliefs should be *less* consequential for act judgments when an act is seen as highly immoral in general. Doing something that is highly immoral in general is bad regardless of whether it violates the actor's own moral standards (that is what it means for an act to be immoral). Performing an act that is *not* immoral in general may be bad if it violates the actor's own moral standards, but acceptable otherwise. This reasoning would predict a larger effect of condition on act judgments when acts are seen as more generally morally acceptable.

In summary, Study 3 was a replication and extension of Study 2, with the following changes: 1) participants only saw a single scenario; 2) participants rated the moral acceptability of the act in general before seeing any information about the actor (rather than afterwards, as in Study 2). Unless noted otherwise, sample size, measures, and analyses were preregistered at AsPredicted.org (https://aspredicted.org/C1J_3SP).

9. Method

9.1. Participants

We recruited 1106 United States residents from Prolific.co in November of 2022. (We increased the sample size for this study to compensate for the lower statistical power of the between-participants design.) After excluding six participants who failed our preregistered attention check as well as seven participants with incomplete data, we were left with 1092 (527 Female, 504 Male, 25 reporting "another" gender, $M_{age} = 37.92$, $SD_{age} = 13.48$). Participants completed the study via a link to a Qualtrics survey in the Prolific interface. Participants were compensated $0.20 for their time (compensation was lower than in previous studies because the current study was much shorter).

9.2. Materials and procedure

Study 3 used the same survey as Study 2, with three modifications. First, participants were randomly assigned to read a single scenario from the 48 used in Studies 1 and 2 (thus they were randomly assigned to one of the control, *not wrong*, or *wrong* conditions in a fully between-participants design). Second, participants rated the general moral acceptability of the act ("How morally acceptable do you, personally, think it is to [act]?" from 1 = *Completely unacceptable* to 7 = *Completely acceptable*) before they read the scenario describing the actor engaging in it, as opposed to after.

Finally, in Study 3 we used a single (preregistered) attention-check question ("How morally acceptable do you, personally, think it is to not pay attention to research surveys? If you are paying attention, please select 'somewhat acceptable'"). Participants who did not select "somewhat acceptable" were excluded.

10. Results

Due to an oversight, we excluded metadesire measures and analyses from our preregistration. We therefore separate the results into preregistered and non-preregistered sections. Note, however, that the metadesire analyses are direct replications of those in Study 2.

10.1. Preregistered analyses

We created an *act wrongness* composite by averaging the two act-focused questions ($\alpha = 0.93$), and a *moral character* composite by averaging the two character-focused questions ($\alpha = 0.90$).

Using OLS regression, we separately modeled act wrongness and moral character composites as a function of condition with fixed effects for scenario (as each participant only saw one scenario, there was no need to adjust standard errors for clustering by participant). These models are shown in Table 4. Compared to the control condition, act wrongness ratings were again significantly higher in the *wrong* condition. Again, there was no difference in act wrongness ratings between the control and *not wrong* conditions. This pattern of results replicates our findings from previous studies. When actors believed their actions to be wrong, they were seen as more wrong. However, when actors did not believe their actions to be wrong, they were not seen as less wrong relative to the control condition.

Moral character ratings also replicated the pattern found in Studies 1 and 2. When actors believed their actions to be wrong, their character was rated more *positively* than in the control condition. When actors did not believe their actions to be wrong, their character was again rated more negatively than in the control condition. Thus, for both act and character ratings, the results of Study 3 replicate the previous two studies.

10.1.1. Moderation by general moral acceptability

We next examined whether the effects of actor beliefs were moderated by participants' own beliefs about the general moral acceptability of the action. To do this, we added each participant's rating of the general acceptability of the action and its interaction with condition as predictors into the regression models described above. Because condition

Table 4
Effects of condition on judgments of act wrongness, actors' moral character, and actor metadesires (Study 3). Models include fixed effects for behavior.

	Act Wrongness			Moral Character			Metadesires		
	Unstandardized *B* [95% CI]	*t*(1,074)	*p*	Unstandardized *B* [95% CI]	*t*(1,074)	*p*	Unstandardized *B* [95% CI]	*t*(1,074)	*p*
Condition									
Baseline – Control	–	–	–	–	–	–	–	–	–
Not Wrong	0.03 [−0.11, 0.17]	0.47	0.64	−0.20 [−0.36, −0.04]	−2.50	0.01	−0.44 [−0.56, −0.31]	−6.83	< 0.001
Wrong	0.16 [0.03, 0.30]	2.32	0.02	0.20 [0.04, 0.35]	2.50	0.01	0.95 [0.83, 1.08]	14.90	< 0.001

was dummy-coded using two dummy variables, models included interactions between general acceptability ratings and each of the dummies (so, two total interaction terms). As preregistered, we focused on the interaction for differences between the "wrong" and "not wrong" conditions, so our statistical test was of the interaction between general acceptability and the dummy contrasting these two conditions (note that this required setting the "wrong" condition as the baseline for these models). These interaction terms were significant both for act wrongness, $B = -0.07$, $t(1{,}071) = -2.62$, $p = .009$, and moral character judgments, $B = 0.09$, $t(1{,}071) = 2.47$, $p = .01$. Interactions are graphed in Fig. 8. These plots show that as the act is seen as less generally morally acceptable, the effect of actor beliefs on character judgments becomes stronger. Conversely, as the act is seen as less generally morally acceptable the effect of actor beliefs on act wrongness judgments becomes weaker. In other words, the effect of actor beliefs on character judgments is strongest for the worst actions, but the inverse is true for act judgments.

10.2. Exploratory analyses

10.2.1. Perceived metadesires

We created a *metadesires* composite by first reverse coding the item, "[Protagonist] wishes s/he **didn't** [act]," and then averaging this item with the other two metadesire questions ($\alpha = 0.81$). Responses scales were numbered such that higher numbers indicates *lower* perceived endorsement of the act by the actor, i.e., more positive perceived metadesires. We report summaries of the analyses of this measure here; full descriptive statistics are available in the Supplemental Material.

Using OLS regression, we first modeled perceived metadesires as a function of condition with fixed effects for scenario. This model showed that compared to the control condition, perceived metadesires were more positive in the "wrong" condition and less positive in the "not wrong" condition (see Table 4).

10.2.2. Mediation by perceived metadesires

As in Study 2, we used the R package *mediation* (Tingley et al., 2014) to estimate the indirect effect of condition on character ratings via perceived metadesires. The average causal mediation effect (ACME) was significant when comparing the *wrong* and control conditions (ACME = 0.26 [95% CI: 0.18, 0.34], $p < .001$) as well as when comparing the *wrong* and *not wrong* conditions (ACME = 0.37 [95% CI: 0.27, 0.48], $p < .001$). Thus, the results of these analyses fully replicate Study 2. For full details of the mediation models from this study, please see the Supplemental Material.

11. General discussion

In three studies, we found that actors' beliefs about their own

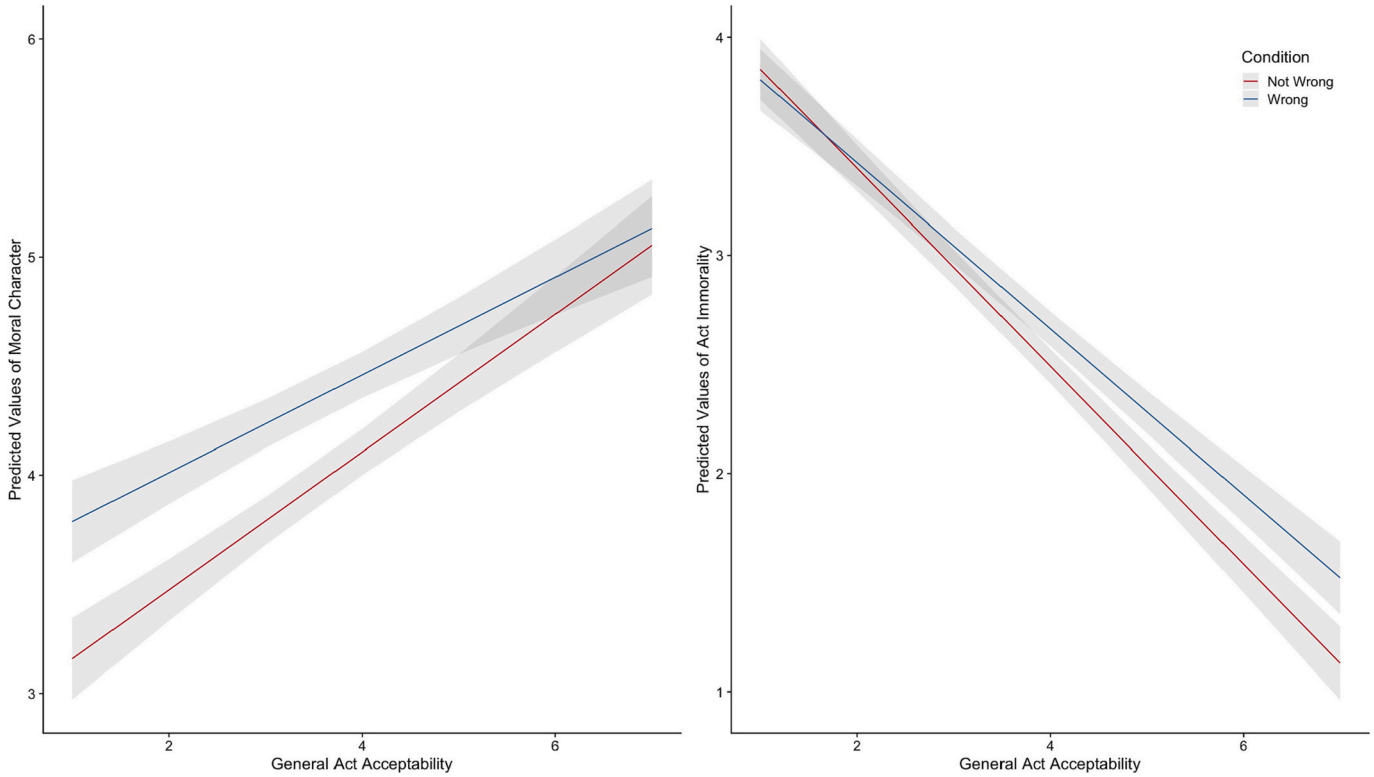

Fig. 8. Predicted ratings of moral character (left panel) and act immorality (right panel) by condition across levels of general moral acceptability of the act. Note. Shaded areas indicate 95% confidence intervals.

everyday immoral behaviors affect both how the acts and the actors are evaluated—albeit in opposite directions. An actor's belief that his or her act is morally wrong causes observers to see the act itself as less morally acceptable; at the same time, it leads to more positive character judgments of the actor. In Study 2, we found that these differences in character judgments were mediated by people's perceptions of the actor's metadesires. Actors who think their behavior is morally wrong are seen as not truly wanting to do what they are doing; this, in turn, leads to more positive evaluations of their character.

Studies 1 and 2 used within-participants designs in which participants saw descriptions of actors who saw their behavior as morally wrong and of actors who did not. In Study 3 we replicated the results of previous studies using a between-participants design in which participants rated only a single actor. This suggests that the current results do not depend on explicitly contrasting actors who see their behavior as wrong with those who do not. Additionally, Study 3 found that the effects of actor beliefs are moderated by participants' prior beliefs about the wrongness of the action in general—but, again, in opposite directions for judgments of actors and acts. The effect of actor beliefs on character judgments is strongest for the most immoral acts whereas the opposite is true for act judgments.

11.1. Perceived metadesires and the positive true self

We found that inferences about actors' metadesires (i.e., inferences about whether they really wanted to act as they did) are responsible for the more positive perceptions of people who think their own behavior is immoral. It is not logically necessary, though, that finding one's behavior to be immoral means that one does not truly want to engage in it. This inference requires some attributional generosity—i.e., the assumption that people's metadesires are normally positive. Indeed, this is what previous research has generally found. Blame is reduced for impulsive negative actions (because people think they are inconsistent with the actor's metadesires) but praise is not reduced for positive impulsive actions (because people think they are *consistent* with the actor's metadesires; Pizarro et al., 2003). More broadly, people believe that others have an underlying "true self" that is morally virtuous and motivates prosocial behavior (De Freitas et al., 2018; Newman, Bloom, & Knobe, 2014; Newman, De Freitas, & Knobe, 2015). The current findings are consistent with this view, but also illustrate a boundary condition. When actors did *not* find their behavior wrong, or even when no information was given, they were judged less favorably (compared to when they were said to find their behavior wrong). Thus, observers are willing to assume positive metadesires when they have some basis for doing so (e.g., an actor's belief that a behavior is wrong) but they are not (or less) willing to do so when there is no basis for it, or when there is in fact evidence to the contrary (e.g., an actor's belief that their negative behavior is not wrong).

In the current studies, participants inferred more positive metadesires for actors who thought their behavior to be wrong even though we specified that the behavior was frequent and ongoing (e.g., "Sam often speeds while driving"). This is somewhat remarkable, especially since for other moral transgressions observers withdraw their attributional generosity when the transgressions are repeated (e.g., Ames & Johar, 2009). It may be that because the immoral behaviors investigated here are relatively minor and common, people find it plausible that others could consistently fall short of their standards despite truly wanting to do better. It may also be that the effect of actor beliefs is even stronger for one-time (vs. repeated) actions. This could be tested more systematically in future research (for example, by varying frequency experimentally).

11.2. Why are act judgments affected by actor beliefs about wrongness?

We found that judgments of the morality of acts showed the opposite pattern as judgments of actors—acts were rated as more wrong and immoral when actors believed them to be morally wrong (compared to the control and *not wrong* conditions, which did not differ significantly in either study). In the Introduction, we outlined two reasons this might be the case. First, actor judgments might change perceiver beliefs about the normative status of those acts in general—e.g., learning that the actor thought it was morally wrong to (for example) jaywalk might lead perceivers to see jaywalking in general as more immoral. We found only weak evidence for this possibility. In Study 1, perceiver ratings of "how morally acceptable do you, personally, think it is to [act]" did slightly differ across conditions; however, including perceiver ratings of behaviors in our models did not change the effects of condition on character or act acceptability. In Study 2, we found no effects of condition on ratings of act acceptability in isolation. Therefore it is unlikely that changes in perceiver beliefs about the normative status of acts in general is responsible for the differences in act ratings we observed.

Second, it might be that people have an explicit or implicit belief that others should not violate their own moral standards, and that doing so is morally wrong (even if the behavior in question is seen as otherwise normatively acceptable). We did not measure these beliefs directly, but we currently see this hypothesis as the most plausible explanation of the pattern of act judgments we observed, particularly because it is compatible with the results of Study 3 (in which we found that the effects of actor beliefs on act judgments are strongest when the acts are seen as normatively acceptable in general). Testing this question more directly is a promising area for future research.

11.3. Future research directions

11.3.1. Signaling positive metadesires but not moral condemnation

In research on moral judgments of hypocrites, Jordan, Sommers, Bloom, and Rand (2017) found that people who publicly espouse a moral standard that they privately violate are judged particularly negatively. However, they also found that "honest hypocrites" (those who publicly condemn a behavior while admitting they engage in it themselves) are judged more positively than traditional hypocrites and equivalently to control transgressors (people who simply engage in the negative behavior without taking a public stand on its acceptability). This might seem to contradict our findings in the current studies, where people who transgressed despite thinking that the behavior was morally wrong were judged more positively than those who simply transgressed. We believe the key distinction that explains the difference between Jordan et al.'s results and ours is that in their studies, hypocrites publicly condemned others for engaging in the behavior in question. As Jordan et al. show, public condemnation is interpreted as a strong signal that someone is unlikely to engage in that behavior themselves; hypocrites therefore are disliked both for engaging in a negative behavior and for falsely signaling (by their public condemnation) that they wouldn't. Honest hypocrites, who explicitly state that they engage in the negative behavior, are not falsely signaling. However, Jordan et al.'s scenarios may have implied to participants that honest hypocrites do condemn others—something that may strike people as unfair coming from a person who engages in the behavior themselves. Thus, honest hypocrites may be penalized for public condemnation, even as they are credited for more positive metadesires.

In contrast, in our studies participants were told that the protagonists thought the behavior was morally wrong but not that they publicly condemned anyone else for engaging in it. This may have allowed protagonists to benefit from more positive perceived metadesires without being penalized for public condemnation. What would happen if actors communicated to a second party both the belief that an act was wrong, and that they themselves engaged in it, but explicitly stated that they did not condemn anyone else for engaging in it? Our prediction is that in this case actors would still be seen as possessing better moral character, but this should be tested empirically in future research.

Relatedly, are people aware of the benefits of signaling that they find their own behavior immoral (and of the pitfalls of appearing to condemn

others)? Given the benefits to moral character judgments, one would expect people to be strongly motivated to communicate moral disapproval of their own negative actions, but whether this is the case (and, if so, whether they are able to avoid backlash from appearing to condemn others) is a question for future research.

11.3.2. Emotions as moral signals

One way in which people might be able to signal negative moral evaluations of their own behavior (and thus, more positive metadesires) is by displaying emotions associated with remorse or moral distress. Indeed, there is some evidence for this possibility. Protagonists who verbally express feeling guilty for accidental harms (e.g., tripping and spilling their coffee on a stranger) are rated as having more positive moral character compared those who do not express guilt (Anderson, Kamtekar, Nichols, & Pizarro, 2021). Likewise, people are blamed less for ambiguously-intentional transgressions (and rated more positively overall) when they display negative facial expressions (vs. when they show neutral or positive expressions; Ames & Johar, 2009). As far as we know, there is no research examining whether people infer metadesires from emotion expressions that accompany behavior, but it seems plausible that they would. The social function of emotions is to signal an individual's internal state to observers (Darwin, 1965; Frank, 1988), and so emotional expressions are likely to be particularly influential signals of underlying dispositions, including morally-relevant metadesires. This is another hypothesis that should be tested directly in future research.

12. Conclusions

Our findings are important first steps in investigating the effects of actor beliefs regarding the moral acceptability of their own actions on observer judgments. In line with predictions made by theories of person-centered morality, we found divergent effects of actor beliefs on ratings of their character and their actions. Knowing how an actor feels about their own behaviors affects not only perceptions of their moral character (positively or negatively) but also of the morality of their action—albeit in opposite directions. These effects seem to be driven by perceiver inferences about actor metadesires. Altogether, the current results suggest that actor moral beliefs are important not only for judgments of their moral character but also for judgments of the relatively benign daily moralized acts in which they engage.

CRediT authorship contribution statement

Stephanie A. Schwartz: Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Yoel Inbar:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision, Visualization.

Data availability

Data and analysis scripts are available at https://researchbox.org/735.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105437.

References

Ames, D. R., & Johar, G. V. (2009). I'll know what you're like when I see how you feel: How and when affective displays influence behavior-based impressions. *Psychological Science, 20*(5), 586–593.

Anderson, R. A., Crockett, M. J., & Pizarro, D. A. (2020). A theory of moral praise. *Trends in Cognitive Sciences, 24*, 694–703.

Anderson, R. A., Kamtekar, R., Nichols, S., & Pizarro, D. A. (2021). "False positive" emotions, responsibility, and moral character. *Cognition, 214*, 104770.

Berman, J. Z., & Small, D. A. (2018). Discipline and desire: On the relative importance of willpower and purity in signaling virtue. *Journal of Experimental Social Psychology, 76*, 220–230.

Critcher, C. R., Helzer, E. G., & Tannenbaum, D. (2020). Moral character evaluation: Testing another's moral-cognitive machinery. *Journal of Experimental Social Psychology, 87*, 103906.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*, 308–315.

Darley, J. M., & Zanna, M. P. (1982). Making moral judgments: Certain culturally transmitted excuses are generally believed to absolve people of blame for harming others. *American Scientist, 70*(5), 515–521.

Darwin, C. (1965). *The expression of emotions in man and animals*. Chicago, IL: University of Chicago Press (Original work published 1872).

De Freitas, J., Sarkissian, H., Newman, G. E., Grossmann, I., De Brigard, F., Luco, A., & Knobe, J. (2018). Consistent belief in a good true self in misanthropes and three interdependent cultures. *Cognitive Science, 42*(Suppl. 1), 134–160. https://doi.org/10.1111/cogs.12505

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878.

Frank, R. (1988). *Passions within reason: The strategic role of the emotions*. New York, NY: Norton.

Frankfurt, H. (1973). Freedom of the will and the concept of a person. *Journal of Philosophy, 68*, 5–20.

Frankfurt, H. (1987). Identification and wholeheartedness. In F. D. Schoeman (Ed.), *Responsibility, character and the emotions* (pp. 27–45). Cambridge, England: Cambridge University Press.

Gendelman, B.. Conflicted omnivores: Meat, morals, and money. *Wharton Research Scholars*. https://repository.upenn.edu/wharton_research_scholars/152.

Goodwin, G. P., Piazza, J., & Rozin, P. (2017). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148.

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science, 6*(8), 859–868.

Greene, J. D. (2009). Fruit flies of the moral mind. In M. Brockman (Ed.), *What's next: Dispatches from the future of science*. New York: Vintage.

Haney, C., Sontag, L., & Costanzo, S. (1994). Deciding to take a life: Capital juries, sentencing instructions, and the jurisprudence of death. *Journal of Social Issues, 50*, 149–176.

Hartman, R., Blakey, W., & Gray, K. (2022). Deconstructing moral character judgments. *Current Opinion in Psychology, 43*, 205–212.

Helzer, E. G., & Critcher, C. R. (2018). What do we evaluate when we evaluate moral character? In K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 99–107). New York: Guilford Press.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309.

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*, 356–368.

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.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114.

Newman, G. E., Bloom, P., & Knobe, J. (2014). Value judgments and the true self. *Personality and Social Psychology Bulletin, 40*(2), 203–216.

Newman, G. E., De Freitas, J., & Knobe, J. (2015). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science, 39*(1), 96–125.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267–272.

Shaver, K. G., & Drown, D. (1986). On causality, responsibility, and self-blame: A theoretical note. *Journal of Personality and Social Psychology, 50*(4), 697.

Shultz, T. R., & Schleifer, M. (1983). Towards a refinement of attribution concepts. In *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 37–62).

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science/ Revue canadienne des sciences du comportement, 13*(3), 238.

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

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). *Mediation: R package for causal mediation analysis*.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*, 72–81.

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

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. Guilford Press.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*, 1251–1263.
Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software, 11*, 1–17.
Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software, 95*, 1–36.