Journal of Experimental Social Psychology 102 (2022) 104371

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Morally questionable actors' meta-perceptions are accurate but overly positive☆

Jeffrey Lees[a,*], Liane Young[b], Adam Waytz[c]

[a] Media Forensics Hub, Clemson University, United States of America
[b] Department of Psychology and Neuroscience, Boston College, United States of America
[c] Kellogg School of Management, Northwestern University, United States of America

ARTICLE INFO

Keywords:
Moral judgment
Meta-perception
Accuracy and bias
Moral motives
Reputation
Preregistered

ABSTRACT

We examine how actors think others perceive their morally questionable behavior (moral meta-perception) across a diverse set of real-world moral violations. Utilizing a novel methodology, we solicit written instances of actors' morally questionable behavior ($N_{\text{total}} = 135$), measure motives and meta-perceptions, then provide these accounts to separate samples of third-party observers ($N_{\text{total}} = 933$), using US convenience and representative samples ($N_{\text{actor-observer pairs}} = 4615$). We find that morally questionable actors can accurately predict how they are perceived, how they are uniquely perceived relative to the average morally questionable actor, and how they are misperceived. Actors who are better at judging the motives of other morally questionable actors also have more accurate meta-perceptions. Yet accuracy is accompanied by two distinct biases: overestimating the positive perceptions others' hold, and believing one's motives are more clearly perceived than they are. These results contribute to a detailed account of the multiple components underlying both accuracy and bias in moral meta-perception.

Navigating the social world requires understanding others' moral preferences and how one's behavior will be perceived by others. These moral meta-perceptions–concerns about how one is judged morally by others–no doubt affect moral decisions (Jordan & Rand, 2020; Rom & Conway, 2018; Vonasch & Sjåstad, 2020), especially for morally questionable acts such as lying to prevent other's emotional distress, or stealing food to feed one's family. Yet less clear is whether individuals have *accurate* insight into how others will evaluate their morally questionable behavior. Can someone who violated a moral norm (e.g., lying to get a job interview) accurately predict others' evaluations of them? Meta-perceptions of stable traits during dyadic interaction tend to be accurate (Carlson & Furr, 2009; Kenny & DePaulo, 1993); yet, judgments of moral behaviors rest not only on perceptions of stable traits, but also on mental state inferences (Schein & Gray, 2018; Young & Tsoi, 2013), motive judgments (Ames & Fiske, 2015; Waytz, Young, & Ginges, 2014), the outcomes of the action itself, such as harm (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014), and the cognitive processes underlying judgment (Conway & Gawronski, 2013; Cushman, 2013).

Past work provides conflicting predictions as to the accuracy of moral meta-perception. In contexts such as political conflict, meta-perception is highly inaccurate as individuals overestimate the negative perceptions of outgroup members (Lees & Cikara, 2020; Moore-Berg, Ankori-Karlinsky, Hameiri, & Bruneau, 2020; Ruggeri et al., 2021; Waytz et al., 2014). More generally, individuals tend to overestimate the strength of others' opinions and expectations (Bursztyn, Gonzalez, & Yanagizawa-Drott, 2018; Prentice & Miller, 1993) and underestimate how positively friends and acquaintances perceive them (Gallrein, Weßels, Carlson, & Leising, 2016). These findings, while not in the domain of moral behavior specifically, suggest morally questionable actors may be overly pessimistic in how they think they will be perceived by others. Conversely, in hypothetical moral dilemmas, which are similar to morally questionable actions in that they involve competing motives and justifications (Bartels and Pizarro, 2011; Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009; Wheeler & Laham, 2016), there is evidence to suggest moral meta-perceptions are relatively accurate. Decision-makers can infer that maximizing moral ends (e.g., preventing as much harm as possible) will lead them to be perceived as intelligent and capable, while respecting prior moral principles (e.g.,

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

[*] Corresponding author at: Wilbur O. and Ann Powers College of Business, Clemson University, 225 Walter T. Cox Blvd., Clemson, SC 29634, United States of America.

E-mail address: jmlees@clemson.edu (J. Lees).

https://doi.org/10.1016/j.jesp.2022.104371
Received 7 February 2022; Received in revised form 26 May 2022; Accepted 27 May 2022
Available online 7 June 2022

never violating someone's rights) will lead them to be perceived as trustworthy and likable (Rom & Conway, 2018; Rom, Weiss, & Conway, 2017), despite observers' inferences of trustworthiness often being miscalibrated (Capraro et al., 2018). Individuals and their acquaintances tend to show agreement on judgments of one's own moral character (Helzer et al., 2014), and narcissists exhibit accurate meta-perceptions regarding their reputations as such (Carlson, Vazire, & Oltmanns, 2011). These disparate findings make it difficult to hypothesize how morally questionable actors' meta-perceptions will compare to how actors are actually perceived in context of real-world moral decisions not induced in a laboratory setting.

To address this question, we integrate research across social perception and moral psychology by adopting a *componential* approach to modeling judgment accuracy. The judgment components we examine are described in Table 1 and adopted from the extant literature on social judgment accuracy (Biesanz, 2010; Funder, 1995; Kenny, 2004; West & Kenny, 2011). This approach has multiple benefits. First, we can disentangle processes which independently contribute to social judgment (e.g., projection of one's own attitudes vs. true accuracy). Second, we can control for confounds which inflate estimates of accuracy in the literature (Wood & Furr, 2016). Third, we can model within-person accuracy, and therefore examine which traits moderate accuracy. Our approach allows us to examine generalized meta-accuracy (Kenny & DePaulo, 1993) and self-other agreement in the context of morally questionable behaviors. For simplicity, the judgment components in Table 1 are described in relation to meta-perception, but all such components apply to observer accuracy as well.

As a conceptual overview of the judgment components, baseline meta-accuracy is the raw linear relationship between how an actor thinks they are perceived and how they are actually perceived. Yet this univariate approach conflates two distinct sources of meta-knowledge: knowledge of how one is uniquely perceived, called distinctive meta-accuracy, and knowledge of how the *average* target is perceived (Wood & Furr, 2016), called normative accuracy. As such, we disaggregate these two components of meta-knowledge. As distinctive meta-accuracy is of greatest theoretical importance, when we examine trait moderators of meta-accuracy we examine if actors higher on those traits have greater distinctive accuracy. Two further facets of meta-accuracy are of theoretical interest: actors' knowledge of how one is *mis*-perceived, called insight (Carlson, 2016), and the extent to which actors overestimate how accurately they will be perceived, called transparency bias (Gilovich, Savitsky, & Medvec, 1998). Lastly, all components of judgment above are conceptualized and operationalized as linear, representing knowledge of the rank-order of others' perceptions *across all* perceptions. We also examine mean point-estimate accuracy of single perceptions, which in this context we call valence biases as all the perceptions are inherently positive or negative. Together these components of judgment provide a theoretically discerning account of generalized meta-accuracy in contexts of morally questionable behavior.

We examine generalized meta-accuracy and self-other agreement across a large set of naturalistic morally questionable behaviors by asking participants to describe in writing a past morally questionable behavior. As the diversity of these written accounts lies in the behaviors described, we ask each participant for self- and meta-perceptions of their *behavior*, rather than global judgments of moral character (e.g., Barranti, Carlson, & Furr, 2016; Helzer et al., 2014). Specifically, in Studies 1–3 we ask participants to rate the moral motives behind their morally questionable behavior (e.g., anger, selfishness, etc.), and in Study 4 we expand the measures to include non-motive attributions (e.g., did the behavior cause harm, etc.). We then take participants' written accounts and provide them to separate samples who rate the behavior on the same measures. This design allows for a direct test of generalized meta-accuracy, self-other agreement, and the additional components of judgment accuracy detailed in Table 1.

Table 1
Detailed breakdown of the theoretical components of judgment we examine, including a description of each component, the specific operationalization, and a model sequence clarifying both the stepwise process by which these models are constructed and whether multiple components are examined within the same model or not.

Judgment Component	Description	Operationalization	Model Sequence
Baseline Meta-Accuracy	Do actors know how they are perceived by others?	Linear relationship (i.e., profile agreement) between meta-perceptions and observer-perceptions (the *true values*)	Model 1
Distinctive Meta-Accuracy	Do actors know how they will be *uniquely (distinctively)* perceived by others, relative to the average actor?	Linear relationship between meta- and observer-perceptions, controlling for the true distribution of attributes in the actor population (the *normative profile*)	Model 2
Normative Accuracy	Do actors know how the average actor is perceived?	Linear relationship between meta-perceptions and the normative profile, in the distinctive accuracy model	Model 2
Meta-Insight	Do actors know how they will be *misperceived*?	Linear relationship between meta- and observer-perceptions, controlling for the actor's true attributes (which is interpretable as the *transparency* of their attributes)	Model 3
Transparency Bias	Do actors overestimate how *transparent* their attributes are to others?	Linear relationship between meta-perceptions and actors' true attributes, in the insight model.	Model 3
Valence Biases	Do actors systematically over or underestimate the extent to which they are attributed a specific attribute?	Mean difference between a specific meta-perception and its corresponding observer perception (true value), across all actors	Model 1
Meta-Accuracy Moderators	Is an actor trait associated with more or less meta-accuracy?	Individual difference measure which moderates the slope of any accuracy coefficient	Models 4+

1. Overview of studies

Table 2 details the Test Sets where we collected written instances of morally questionable behaviors (“something bad for a good reason”),

Table 2
Breakdown of the samples for each Actor Test Set and Observer Study, including sample sizes, sample characteristics, and the number of judgment items. Horizontal correspondence represents which Test Sets each Observer Sample rated. Study 3 corresponds with both Test Sets because Observers in Study 3 rated Test Set A, but were themselves the Actors in Test Set B.

Actor Test Sets	Observer Studies
Test Set A *N* = 40, convenience sample 13 stories chosen 6 motive/judgment items	Study 1, N = 318, convenience sample Study 2, N = 121, convenience sample Study 3, N = 230, US representative sample
Test Set B N = 230, US representative sample 122 stories chosen 27 motive/judgment items	Study 4, *N* = 256, US representative sample

and accompanying self- and meta-perceptions (henceforth referred to as “stories” written by “actors”), and the Observer studies where we showed a separate sample of participants a selected subset of stories from the corresponding Test Sets.

Observers in Studies 1 and 2 read the stories from Test Set A. The actors in Test Set B were the *same participants* as the observers in Study 3. These participants first provided their written stories (Test Set B), then read and judged the stories from Test Set A (Study 3). Observers in Study 4 read the stories from Test Set B.

2. Open Science

Data collection and analyses for Study 2 were preregistered, (https://osf.io/vcxb4), as were Test Set2/Study 3 (https://osf.io/bgv85), and Study 4 (https://osf.io/kzybe). All materials, data, and analysis code for all studies are available on the OSF (https://osf.io/k6hms). We report all measures, and any deviations from the preregistrations are noted. Full disclosure to participants of our research question and intent was utilized across all studies.

The preregistrations included every confirmatory and exploratory analysis we intended to conduct, however this means we conducted myriad ancillary analyses which are impractical to include in the main text. As such, every preregistered analysis is reproduced in the Open Science Framework materials (https://osf.io/k6hms). None of the analyses therein contradict the results presented here in the main text.

Complete details and results of all the linear mixed-effect models reported in summary below can be found in Section 3 of the Supplemental Materials, including regression tables, random effect variance components, and variance-covariance matrices for all models.

3. Test Set A

3.1. Method

3.1.1. Participants and design

Test Set A solicited stories from participants regarding a previously committed unethical behavior. Forty individuals (Mean$_{age}$ = 31.9 years, 24 Men, 16 Women) were recruited for a short online survey using Amazon's Mechanical Turk platform, including the TurkPrime features (Litman, Robinson, & Abberbock, 2017). The study was advertised as taking 3–4 min to complete (Median$_{duration}$ = 170 s) and participants were paid $0.60 upon completion. The main purpose of Test Set A was to capture approximately a dozen stories of sufficient quality that could be used as stimuli in future studies. Because we could not anticipate either the quality of stories or how many participants would consent to allowing us to share their stories, we decided a priori to collect 40 participants in the hope of obtaining at least twelve stories of reasonable quality and variety.

3.1.2. Procedure

Participants were informed they would be asked for potentially sensitive information and then received a prompt to write about an instance of morally questionable behavior that read:

> Sometimes people have to ‘do bad things for good reason.’ Examples could include stealing food to feed your family, harming someone in order to protect yourself, or lying to friends or family in order to prevent worse conflict. In service of a greater good, we sometimes have to commit lesser evils. Please describe, in 2–4 sentences, a time when you did something ‘bad,’ but for good reason, or because you felt it was necessary. After providing your story you'll be asked a few short questions about your reasoning behind your behavior. Please do not include any identifying information in your story.

This language of “something bad for a good reason” was chosen in the hope of soliciting stories that represented morally “gray” behaviors with complex and competing motives.

After writing, participants indicated their actual motives and meta-motives (order of sections randomized) for the behaviors they described. Participants were not aware of the specific items/motives they would be asked about when they were writing their stories, as to not bias what or how participants write about their behavior. Nor could participants return to their written story to edit it once they began responding to the motive items. The actual motive instructions read “when engaging in the behavior you just described, how much were you motivated by the following,” and the meta-motive instructions read “Imagine someone else—the average person—reads about the behavior your described and has to rate how much they think you were motivated by the following factors (below). Please indicate how you think THEY would rate your motivations.” Response options were: selflessness, compassion, loyalty, self-interest, anger, and conflict avoidance (single items, all rated from 1=“Not at all” to 7 = “Completely”). These six motive items were selected for their generality and potential applicability to behaviors participants might describe. They intentionally capture both positively and negatively valenced motives, emotional and non-emotional motives, along with self- and other-focused motives (note that in Test Set B the list of motives expands greatly).

After rating their motives and meta-motives participants indicated whether or not we could use their stories in future studies (two withheld consent and are not included in any studies). Finally, participants completed demographic questions and the survey ended.

4. Results

Table 3 provides the bivariate correlations between participants' actual motives and meta-motives across the 38 stories. The high correlations between each self-reported motive and its corresponding meta-motives (*r*s from 0.68 to 0.88, bolded) suggested that actors did not expect observers' perceptions to deviate significantly from their self-reported motives. This potentially represented a transparency bias (Gilovich et al., 1998), where individuals believe their thoughts, feelings and emotions are more transparent to observers than they actually are.

Of the 38 stories we selected 13 to serve as stimuli for future studies. These were selected for their quality (i.e. they were relatively well written, comprehensible, and believable), and diversity (i.e. they represent a wide set of behaviors and contexts). The 13 we chose are accessible on the OSF (https://osf.io/k6hms). Below are brief summaries of the content of each story. Story #1: Called in sick (falsely) to go on their child's field trip. Story #2: Shoplifted food during a brief period of homelessness. Story #3: Dropped out of college and didn't tell parents. Story #4: Searched sister's browser history to learn if she was self-harming and got her help. Story #5: Stole $5 from their brother to give to a homeless man. Story #6: Put fake references on job applications. Story #7: Lied to a recently-single friend about having a new boyfriend. Story #8: Argued/almost fought shopkeeper who swindled brother. Story #9: Stole food for their cat when financially broke. Story #10: Didn't tell father about cancer diagnosis. Story #11: Lied to friend about having transgender daughter. Story #12: Stole food for an impoverished friend. Story #13: Broke suicidal friend's promise to remain silent in order to get them help.

5. Study 1

Study 1 took the 13 stories from Test Set A and used them as stimuli for participants to rate, along with the same six motive items from Test Set A. We then directly tested for meta-accuracy by comparing Test Set A's meta-perceptions to Study 1's observer perceptions.

5.1. Method

5.1.1. Participants and design

We decided a priori to collect 25 responses per story, requiring a total of 325 participants. 326 participants were recruited via Mturk using

Table 3

Pearson correlations between actors' own motives and meta-motives. $N = 38$. Bolded values are the specific congruence between each motive and its corresponding meta-motive. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	Selfless.	Compass.	Loyalty	Self Interest	Anger	Conflict Avoid.	Meta Selfless.	Meta Compass.	Meta Loyalty	Meta Self Interest	Meta Anger	Meta Conflict Avoid
Selflessness												
Compassion	0.64***											
Loyalty	0.29	0.51**										
Self-Interest	−0.44**	−0.56***	−0.29									
Anger	−0.00	−0.11	0.03	0.33*								
Conflict Avoidance	−0.18	−0.43**	−0.37*	0.17	−0.01							
Meta Selfless	**0.88*****	0.63***	0.36*	−0.50**	−0.20	−0.11						
Meta Compassion	0.77***	**0.71*****	0.43**	−0.65***	−0.18	−0.12	0.83***					
Meta Loyalty	0.55***	0.56***	**0.68*****	−0.45**	−0.14	−0.31	0.60***	0.74***				
Meta Self Interest	−0.43**	−0.67***	−0.39*	**0.76*****	0.23	0.40*	−0.52***	−0.60***	−0.48**			
Meta Anger	−0.05	−0.02	0.08	0.39*	**0.79*****	−0.04	−0.18	−0.14	−0.15	0.24		
Meta Conflict Avoid	−0.06	−0.29	−0.29	0.20	0.05	**0.88*****	−0.03	−0.06	−0.24	0.49**	0.08	

Computed correlation used pearson-method with pairwise-deletion.

TurkPrime, and eight responses were removed due to duplicate IP addresses, leaving a final $N = 318$ (Mean$_{\text{age}}$ = 35.6 years, 175 Men, 143 Women). The study was advertised as taking 1–2 min (Median$_{\text{duration}}$ = 66 s) and participants were paid $0.25 upon completion.

5.1.2. *Procedure*

After providing informed consent, participants were informed that they would be reading and judging a story of morally questionable behavior written by another participant, and the language of these instructions matched the language from the writing prompt in Test Set A: "Sometimes people have to 'do bad things for good reason.' Examples could include stealing food to feed your family, harming someone in order to protect yourself, or lying to friends or family in order to prevent worse conflict. In service of a greater good, we sometimes have to commit lesser evils. You are going to read a story of when someone did something 'bad,' but for good reason, or because they felt it was necessary. The story was written by another participant in this study. After reading the story you'll be asked a few short questions about your thoughts regarding the story."

On the following page participants were randomly assigned to read one of the thirteen stories, unaltered from what was originally written by participants in Test Set A. On the same page as the story participants were asked to rate how much they believed the story's author was motivated by the same six motives on the exact same scale as Test Set A. After reading a single story participants responded to basic demographic questions and the survey ended.

5.1.3. *Analyses*

All our analyses were adapted from the social accuracy model (SAM: Biesanz, 2010), see Section1 of the Supplementary Materials for detailed modeling information, including data centering details. We used true-mean (i.e., the accuracy criterion) centered judgment variables within target ("person-centering," Furr & Funder, 2004), after combining the observer data in Study 1 with actors' meta- and actual motives collected in Test Set A. As such, actors' meta-perceptions were true-mean centered on the average observer perception toward the specific actor, observers' attributions were true-mean centered on the average actual motives reported by actor, and actors' actual motives were mean-centered within actor. Standardized regression coefficients are provided and were derived using "within-group" standardization (Hamaker & Muthén, 2020; Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016), where all judgments are divided by the within-actor standard deviation, rather than the grand standard deviation used in traditional standardization. Statistical details on the centering and standardization method can be found in Section 1 of the Supplemental Materials. Complete regression tables, including variance-covariance matrices, for all the models reported below can be found in Section 3 of the Supplemental Materials.

For both actor meta-accuracy and observer judgment accuracy, we tested three main models utilizing linear mixed-effects modeling, adapted from the SAM framework. Note that observer accuracy was operationalized as profile-agreement, the extent to which observers' judgments of actors' motives related to actors' self-reported motives (the true value). The first models were baseline accuracy models, which tested for a linear relationship between the estimated and true values in actor meta-perceptions and observer motive attributions. The baseline accuracy models should be interpreted with caution as they do not control for potential confounds (controlled for in subsequent models) of judgment accuracy which may inflate accuracy estimates.

The second models tested for *distinctive accuracy* (Biesanz, 2010; Furr, 2008), which tested the linear relationship between the estimated and true values while simultaneously testing for the linear relationship between the estimated value and the *normative profile* (sometimes referred to as the stereotypic profile). Conceptually, it is possible that observed accuracy is due, in part or in whole, to an observer's reliance on expectations/knowledge regarding the average distribution of motives in the population, rather than the observer's consideration of the motives at play in the specific story they were judging. Across all studies we operationalize the normative profile as the distribution of the motive means across the actor sample, given that the relevant normative profile for participants is specific to our paradigm.

The third models tested for *insight* in meta- and observer judgments (Carlson, 2016). For actors, insight was operationalized as the linear relationship between their meta-perceptions and observers' motive attributions while controlling for actors' actual motives, the latter relationship interpreted as *transparency bias* in actors' motives as it would indicate that actors think their true motives are perceived when they are not in truth. Conceptually, actor insight represented the ability of an actor to accurately predict how perceptions of their motives *differed* from their true motives (i.e., how they're misperceived). For observers, insight was operationalized as the linear relationship between their motive attributions and actors' true motive while controlling for actors' meta-perceptions, which here was interpreted as *opaqueness* in actors' motives. Conceptually, observer insight represented the ability of an observer to accurately perceive the extent to which an actor's true motives *differed* from how the actor thought they would be perceived.

In constructing our linear mixed-effects models we adopted a maximal random structures approach (Barr, Levy, Scheepers, & Tily, 2013), with random slopes and intercepts for all linear predictors across all random effects (unless a model was singular, in which case we removed the random slope with the smallest standard deviation). Building upon the SAM our models were constructed such that the estimates provided by participants (i.e. the meta-perceptions, observer motive judgments) were modeled as the dependent variables, and the true values were modeled as predictors. In predicting actor meta-perceptive accuracy there was a single random effect for actor, and in predicting observer accuracy there were two crossed random effects, one for actor and one for observer (see Section 2 of the Supplemental Materials for information on modeling the interaction of actor and observer as separate random intercepts). All models were estimated using restricted maximum likelihood estimation and Welch–Satterthwaite degrees of freedom approximation through the *lmerTest* R package (Kuznetsova, Brockhoff, & Christensen, 2017).

5.1.4. Sensitivity analyses

We conducted Monte Carlo simulations, using the *simr* R package (Green & MacLeod, 2016), of the models used to test our key hypotheses to examine the sensitivity with which they would be able to observe effect sizes at and below the observed estimate. We examined the sensitivity of the distinctive meta-accuracy and distinctive observer accuracy models, as these estimates are of primary theoretical interest and are comparable in structure to the other analyses. Sample size for the simulations was $N_{actors} = 13$ and $N_{observers} = 318$ (1899 actor-observer judgments). We estimated the statistical power of the observed distinctive meta-accuracy effect size ($b = 0.23$, see below), and the sensitivity to detect smaller estimates (0.18 and 0.13), based on 500 simulations for each estimate, with alpha = 0.05. The observed estimate of $b = 0.23$ was powered at 96%, 95% CI = [0.94, 0.98], while $b = 0.18$ was observable with 89% power, 95% CI = [0.85, 0.91], and $b = 0.13$ observable with 63% power, 95% CI = [0.59, 0.67]. We estimated the statistical power of the observed distinctive observer-accuracy effect size ($b = 0.31$, see below), and the sensitivity to detect smaller estimates (0.25 and 0.18), based on 500 simulations for each estimate. The observed estimate of $b = 0.31$ was powered at 99.8%, 95% CI = [0.99, 1.00], while $b = 0.25$ was observable with 98% power, 95% CI = [0.97, 0.99], and $b = 0.18$ was observable with 83% power, 95% CI = [0.80, 0.87].

5.2. Results

5.2.1. Meta-perception accuracy

The baseline meta-accuracy model found evidence for meta-accuracy, $b = 0.41$, 95% CI = [0.29, 0.54], $B = 0.43$, $t(12.15) = 7.12$, $p < 0.001$. Actors were able to predict with some accuracy the motive attributions observers made toward them. The distinctive accuracy model, which controlled for the *average* distribution of motives (the normative profile), found evidence for distinctive accuracy, $b = 0.23$, 95% CI = [0.11, 0.35], $B = 0.25$, $t(12.04) = 4.28$, $p = 0.001$, meaning actors exhibited meta-accuracy in judging how they're *uniquely* perceived relative to the average morally questionable actor. The normative profile was also associated with meta-perceptions, $b = 0.83$, 95% CI = [0.34, 1.32], $B = 0.39$, $t(12.04) = 3.71$, $p = 0.003$. Lastly, the meta-insight model, which controlled for actors' true motives, found evidence of meta-insight, $b = 0.04$, 95% CI = [0.02, 0.06], $B = 0.04$, $t(1874.63) = 4.62$, $p < 0.001$, and that meta-perceptions were associated with actors' true motives, $b = 0.79$, 95% CI = [0.59, 1.00], $B = 0.83$, $t(11.73) = 8.59$, $p < 0.001$. Interpretively, this suggested that actors possess, to a small degree, knowledge about how they would be misperceived, and that they displayed a transparency bias where they assumed that observers' judgments would track more closely with their actual motives than they did in truth.

5.2.2. Observer accuracy

The observer base accuracy model found that observers were able to accurately assess the self-reported motives of the actors in the story they read, $b = 0.45$, 95% CI = [0.36, 0.55], $B = 0.42$, $t(6.48) = 11.72$, $p < 0.001$. We also found evidence for distinctive accuracy, $b = 0.31$, 95% CI = [0.18, 0.44], $B = 0.30$, $t(10.46) = 5.44$, $p < 0.001$, suggesting observer's judgments tracked with the distinctness of the actor they are perceiving while also tracking with the normative profile, $b = 0.41$, 95% CI = [0.20, 0.62], $B = 0.20$, $t(13.02) = 4.24$, $p = 0.001$. Lastly, the insight model found evidence for accuracy, $b = 0.14$, 95% CI = [0.03, 0.24], $B = 0.11$, $t(83.60) = 2.48$, $p = 0.015$, in addition to an association between observers' judgments and actors' meta-perceptions, $b = 0.36$, 95% CI = [0.21, 0.50], $B = 0.33$, $t(15.31) = 5.34$, $p < 0.001$. These findings suggest that observers were both able to perceive actor's motives accurately but that actors' motives were also partially "opaque." Operationally, opaqueness was exhibited when observers' judgments were associated with actors' *meta-perceptions* above actors' true motives, meaning observers perceived some motives inaccurately and in a manner that actors anticipated.

6. Study 2

Study 2 sought to directly replicate the findings of Study 1 while also expanding the survey to include observer trait measures and secondary attributions toward the stories. Study 2 was preregistered (https://osf.io/vcxb4).

6.1. Methods

6.1.1. Participants and design

We recruited 125 participants to take an online survey on Mturk, using TurkPrime. Four participants failed the comprehension check and as such the final $N = 121$ ($\text{Mean}_{age} = 34.3$ years, 81 Men, 39 Women, 1 Non-binary individual). We preregistered a sample size of 125 to detect the small level of meta-accuracy among actors from Study 1. In Study 1 we gathered 25 responses per story, and in Study 2 we utilized repeated measures and on average obtained approximately 47 responses per story. The study was advertised as taking 15 min ($\text{Median}_{duration} = 11$ min, 54 s) and participants were paid $2.25 upon completion.

6.1.2. Procedure

The design of Study 2 was identical to that of Study 1 except for two main features: Study 2 introduced repeated measures where participants read five, instead of one, story (randomized), and Study 2 included additional measures within story and participant trait measures. Otherwise the procedure was identical to that of Study 1 (same instructions, same motive items, etc.).

For each of the five stories observers read, they rated the actors on the six motive items from Test Set A along with a series of new judgment items about the actors, and several trait measures. Both the new judgment items and the trait measures were included so that we could test whether they moderated observer accuracy (were associated with greater or less accuracy). We included a three-item measure of empathy for the actor ("How much do you feel compassion for /sympathy for/ moved by the story author?"), a three-item measure of perceived immorality ("is this behavior immoral/ethical(-)/right or wrong?"), a single item for perceived similarity to the actor ("How similar/dissimilar is the author of this story to you"), a single item for perceived trustworthiness of the actor ("How trustworthy is the story author?"), and a single item for perceived typicality of the behavior ("How common is the story author's behavior?"). After reading and rating the five stories participants completed several trait measures: the perspective taking ($\alpha = 0.89$) and empathic-concern ($\alpha = 0.94$) subscales of the Interpersonal Reactivity Index (Davis, 1983), trait Machiavellianism ($\alpha = 0.87$) (Dahling, Whitaker, & Levy, 2009), and cognitive ability as measured by twelve Raven's Progressive Matrices ($\text{Mean}_{correct} = 5.42$, SD = 2.78) (Raven, 2000).

After completing these measures participants completed a comprehension check that they asked “Earlier in the survey you read several stories. What was the nature of those stories, collectively?” The correct answer, among seven options, was “moral behavior” (incorrect answers included unrelated topics such as “culinary preferences” and “consumer purchasing habits”). The four participants who failed the comprehension check were excluded from all analyses. Afterward participants answered basic demographic questions, including a single item measure of political orientation, and the survey ended.

6.1.3. *Analyses*

In our efforts to directly replicate the findings from Study 1 in Study 2 we began by performing the exact analyses that we performed in Study 1, adapted from the SAM (Biesanz, 2010). As such we examined accuracy, distinctive accuracy, and insight, for both actors and observers. See Section 1 of the Supplementary Materials for detailed modeling information. In addition to reperforming these exact analyses we further explored the role of our new measures in moderating observer accuracy. To do so we entered the new measures separately as interactions with the true-value (actors' true motives) in the distinctive accuracy models. This allowed us to examine whether distinctive accuracy was related to other judgments and/or traits. Per the preregistration, all direct replications of the findings from Study 1 were confirmatory analyses (except the “insight” models, which were not preregistered), while all examinations of accuracy-moderation by new variables were exploratory analyses. All trait measures were mean-centered and all judgment measures were true-mean centered within target. Complete regression tables, including variance-covariance matrices, for all the models reported below can be found in Section 3 of the Supplemental Materials.

6.1.4. *Sensitivity analyses*

We conducted Monte Carlo simulations, using the *simr* R package (Green & MacLeod, 2016), of the models used to test our key hypotheses to examine the sensitivity with which they would be able to observe effect sizes at and below the observed estimate. We examined the sensitivity of the distinctive meta-accuracy and distinctive observer accuracy models. Sample size for the simulations was $N_{actors} = 13$ and $N_{observers} = 121$ (3626 actor-observer judgments). We estimated the statistical power of the observed distinctive meta-accuracy effect size ($b = 0.26$, see below), and the sensitivity to detect smaller estimates (0.20 and 0.10), based on 500 simulations for each estimate, with alpha = 0.05. The observed estimate of $b = 0.26$ was powered at 97%, 95% CI = [0.95, 0.98], while $b = 0.20$ was observable with 81% power, 95% CI = [0.77, 0.84], and $b = 0.10$ observable with 30% power, 95% CI = [0.26, 0.34]. We estimated the statistical power of the observed distinctive observer-accuracy effect size ($b = 0.37$, see below), and the sensitivity to detect smaller estimates (0.25 and 0.15), based on 500 simulations for each estimate. The observed estimate of $b = 0.37$ was powered at 96%, 95% CI = [0.94, 0.97], while $b = 0.25$ was observable with 74% power, 95% CI = [0.70, 0.78], and $b = 0.15$ was observable with 33% power, 95% CI = [0.29, 0.37].

6.2. *Results*

6.2.1. *Meta-perception accuracy*

The baseline meta-accuracy model found evidence for meta-accuracy, $b = 0.42$, 95% CI = [0.30, 0.55], $B = 0.47$, $t(12.06) = 7.20$, $p < 0.001$, meaning that actors could predict with some accuracy the motive attributions observers made toward them. The distinctive accuracy model found evidence for distinctive accuracy, $b = 0.26$, 95% CI = [0.12, 0.41], $B = 0.30$, $t(12.05) = 4.04$, $p = 0.002$, and normative accuracy, $b = 0.79$, 95% CI = [0.30, 1.29], $B = 0.37$, $t(12.03) = 3.51$, $p = 0.004$. We found evidence of meta-insight, $b = 0.03$, 95% CI = [0.005, 0.05], $B = 0.03$, $t(11.89) = 2.62$, $p = 0.023$, and transparency bias, $b = 0.79$, 95% CI = [0.59, 1.00], $B = 0.83$, $t(11.86) = 8.54$, $p < 0.001$. These results directly replicated the findings from Study 1.

6.2.2. *Observer accuracy*

Replicating the findings from Study 1, the observer accuracy models found that observers were able to accurately assess the self-reported motives of the actors in the stories they read across baseline accuracy, $b = 0.51$, 95% CI = [0.35, 0.66], $B = 0.47$, $t(11.51) = 7.06$, $p < 0.001$, and distinctive accuracy, $b = 0.37$, 95% CI = [0.17, 0.58], $B = 0.35$, $t(11.39) = 3.97$, p = 0.002, along with displaying normative accuracy, $b = 0.37$, 95% CI = [0.16, 0.59], $B = 0.18$, $t(11.47) = 3.39$, $p = 0.006$. However we did not replicate the finding that observers perceived actors' true motives beyond how actors' thought they would be perceived (observer insight), $b = 0.04$, 95% CI = [−0.41, 0.48], $B = 0.04$, $t(10.45) = 0.17$, $p = 0.865$, and found only a relationship between observers' judgments and how actors' thought they would be perceived (opaqueness), $b = 0.53$, 95% CI = [0.07, 0.98], $B = 0.44$, $t(10.73) = 2.53$, $p = 0.028$. The significant distinctive accuracy finding but insignificant insight finding suggests that observers can assess the unique motives of actors, but that actors' most concealed motives are still not fully perceived by observers.

6.2.3. *Moderators of observer accuracy*

To examine moderators of observer accuracy we first examined traits as potential moderators. We separately entered each trait as an interaction with actors' true motives in predicting observers' attributions while controlling for the normative profile (the distinctive accuracy model). A significant interaction would indicate that the trait moderates observer-accuracy, with a positive (negative) coefficient meaning individuals higher on the trait are significantly more (less) accurate. We found that cognitive ability, $b = 0.01$, 95% CI = [0.002, 0.03], $B = 0.04$, $t(121.08) = 2.29$, $p = 0.024$, and empathic-concern, $b = 0.04$, 95% CI = [0.01, 0.06], $B = 0.05$, $t(122.81) = 3.08$, $p = 0.003$, positively moderated distinctive accuracy, meaning observers higher on those traits were more accurate in their motive judgments. Conversely, Machiavellianism negatively moderated distinctive accuracy, $b = -0.04$, 95% CI = [−0.08, −0.01], $B = -0.04$, $t(118.57) = -2.61$, $p = 0.010$, meaning those high in Machiavellianism (those who self-reporting being manipulative, distrusting, and status seeking) were significantly less accurate in judging the unique motives of morally questionable actors. Trait-perspective taking did not moderate observers' distinctive accuracy, $b = 0.01$, 95% CI = [−0.02, 0.04], $B = 0.01$, $t(118.02) = 0.76$, $p = 0.449$.

For non-trait moderators of accuracy, we found that perceived trustworthiness of the actor positively moderated observer distinctive accuracy, $b = 0.02$, 95% CI = [0.002, 0.04], $B = 0.03$, $t(1207.29) = 2.13$, $p = 0.034$, along with perceived similarity, $b = 0.02$, 95% CI = [0.003, 0.04], $B = 0.04$, $t(681.30) = 2.26$, $p = 0.024$, meaning that observers who perceived the actors as more trustworthy and similar to themselves were also more accurate at judging the unique motives of the actors. Because both judgment accuracy and perceptions of trustworthiness and similarity were all within-target repeated measures, and we centered our judgment data within-target to orthogonalize within and between participant variance, we can infer that the positive moderating effects of perceived trustworthiness and similarity are *within-observers* effects. As a single observer's attributions of trustworthiness and similarity increase across targets, their accuracy in judging the unique motives of their target increase. This also means that the positive relationship between attributed trustworthiness/similarity and judgment accuracy cannot be explained by the possibility that participants who are more trusting also happen to be more accurate (for example), or are also high in unobserved third variables which predict judgments accuracy. None of the other observer-judgment items moderated observer accuracy, namely empathy for the target, $b = 0.00$, 95% CI = [−0.02, 0.02], $B = 0.01$, $t(571.52) = 0.40$, $p = 0.692$, perceived immorality, $b = -0.01$, 95% CI = [−0.03, 0.01], $B = -0.02$, $t(1512.53) = -0.67$, $p = 0.506$, and perceived typicality of the behavior, $b = 0.00$, 95% CI = [−0.02, 0.02], $B = 0.00$, $t(702.54) = 0.07$, $p = 0.942$.

7. Test Set B

Test Set B collected 230 stories and we chose 122 to use as stimuli. It also expanded the number of judgment items in the survey from six to 27. Test Set B was preregistered (https://osf.io/bgv85).

7.1. Methods

7.1.1. Participants and design

230 participants (Mean$_{\text{age}}$ = 40 years, 115 Men, 113 Women, 2 Other-Gender) were recruited to participate in a 15–18 min long study through Qualtrics survey panels. A quota-matching system was utilized to guarantee the final sample would be representative of the general US population, with quotas set to census distributions along the following demographic characteristics: age, gender, ethnicity, education, and income. The goal was to collect 130 stories of sufficient quality to use as stimuli, as 130 actors would give us 80% power to detect the small moderation effect of empathic-concern on distinctive accuracy we observed in Study 2. However, after collecting 177 responses we found that only 104 stories meet our criteria for coherence and specificity. As such, we performed a second collection of data to increase the sample size (this was also preregistered: https://osf.io/4dywx). Both samplings were combined, totaling 230 stories, and we deemed 122 to meet our criteria for story quality.

7.1.2. Procedure

After providing informed consent and responding to demographic questions (for the quotas), participants read the exact same prompt from Test Set A which elicited from them their written accounts of morally questionable behavior. Participants then immediately completed a comprehension check and if they failed they were prevented from continuing the survey. The following page displayed the exact text of the story participants had written across the top of the screen and asked them for their actual motive and meta-motive (counterbalanced).

In Test Set B we greatly expanded upon the measures participants provided at the self/meta-level, while maintaining the breadth of positively and negatively valenced motives, emotional and non-emotional motives, and self- and other-focused motives, as to ensure generality and potential applicability to any and all behaviors participants may describe. The six motives used in Test Set A were expanded to sixteen in Test Set B. Additionally, we added eleven moral judgments items unrelated to motives, to expand the theoretical scope of moral attributions individuals make of themselves and of others' moral behaviors. As such, Test Set B includes a larger and more generalizable set of moral attributions compared to Test Set A.

The sixteen motive items used were based on existing theories in moral psychology on the motivational antecedents of moral behavior. We chose to construct items based on multiple, and often conflicting, theories of moral motives because in our paradigm participants were able to describe *any* type of moral violation yet were only able to express their motives via the items we provided. As such, diversity and breadth in potential motives for participants to report was the central methodological motivation behind item construct. The sixteen motive items included "Inhibition", "Avoiding Conflict", "Independence", and "Helping Others", reflecting approach and avoidance motives (Janoff-Bulman, Sheikh, & Baldacci, 2008); "Supporting a Friend", "Loyalty", "Fairness", and "Punishing Others", reflecting relationship regulation motives (Rai & Fiske, 2011); "Anger", "Guilt", "Compassion", and "Pride", reflecting moral emotions (Cameron, Lindquist, & Gray, 2015; Tangney, Stuewig, & Mashek, 2007; Teper, Zhong, & Inzlicht, 2015); "Selflessness" and "Self-Interests" (Barasch, Levine, Berman, & Small, 2014; Batson, Klein, Highberger, & Shaw, 1995); and "Duty" and "Obligation" (Baron, Ritov, & Greene, 2011; Gerstenberg et al., 2018). All motive items were single items rated from 1="Not at all" to 7 = "Completely."

In addition to these sixteen motive items asked at the self and meta level, we also asked at the self and meta level perceived immorality of the behavior (3-items, $\alpha_{\text{meta}} = 0.81$, $\alpha_{\text{self}} = 0.85$) using the same language as the items in Study 2, whether the behavior pertained to "moral" and "non-moral considerations", actor trustworthiness, typicality/intentionality/harmfulness of the behavior, and whether the behavior "violated" and "fulfilled an obligation". Whether participants received the meta-items or self-items first was counterbalanced, as was the order of the items within each block. These broad, non-motive judgment items were added so that actors could express relevant perceptions of their own behavior beyond the domain of motivation, again given that actors were able to describe *any* type of moral violation with the paradigm.

On the following page we explicitly asked participants for permission to use their written story as stimuli in future research, using the same language as Test Set A. Of the 230 participants 206 granted permission and 24 withheld it. We do not report any information regarding those 24 stories where permission was withheld, nor were they included in any analyses of the written stories.

Next, participants completed the trait measures in counterbalanced order. Participants responded to the empathic-concern ($\alpha = 0.77$) and perspective-taking ($\alpha = 0.77$) subscales of the IRI (Davis, 1983), trait Machiavellianism ($\alpha = 0.91$) (Dahling et al., 2009), propensity to engage in unethical workplace behaviors ($\alpha = 0.95$) (Chen & Tang, 2006), and twelve Raven's Matrices as a measure of cognitive ability (Mean$_{\text{Correct}}$ = 3.4, SD$_{\text{Correct}}$ = 2.7) (Raven, 2000).

Participants then shifted to being observers of the stories from Test Set A, where they responded to a random set of six of the thirteen stories. See section on Study 3 for details. Afterwards participants provided their age, responded to a single-item measure of political orientation, and lastly were given an open text box to provide any further details regarding their story they felt the researchers ought to know.

7.2. Results

Similar to Test Set A, there were high bivariate correlations between actors' self judgments (perceptions and motives) and their meta-perceptions. Across the sixteen motive items and eleven judgment items the average bivariate correlation between self- and meta-items for the 122 chosen stories was $r = 0.65$ (highest $r = 0.80$ for helping motive, lowest $r = 0.39$ for non-moral considerations, for all correlations $p < 0.001$). Overall, these high correlations suggest that actors believed they will be perceived relatively accurately by observers, further providing evidence for a transparency bias (Gilovich et al., 1998).

In considering which of the 206 stories we would use as stimuli in subsequent studies (here Study 4), where they would be given to a new set of third-party observers, we based our decisions on two criteria: coherence and specificity of the story. We did not include/exclude stories based on the nature of the behavior. For example, if the behavior described was abjectly immoral and seemed to not reflect any "good" motives, per our prompt to participants, it was still a candidate for inclusion. As long as the story was coherent enough for a reasonable person to understand and specific enough that a reasonable person could comprehend the basics of the context the story was included. For example, a story we did not include because of a lack of specificity was "I had to lie to someone in order to protect their feelings," whereas we did choose to use the story "I lied to my spouse so we wouldn't have a fight about something I ate." This latter story contains specific information about the lie and context that the former story lacks. Of the 206 stories we deemed 122 to meet our criteria.

Using the categories of domain-general moral violations from Powell and Horne (2017), we rated each story for the presence of a given moral violation for the purpose of providing a qualitative overview of the types of behaviors in Test Set 2's stories. Table 4 contains the percentage of the 122 final stories which included each moral violation. Qualitatively, the "Other" behaviors varied widely, with some common behaviors including the sale of illicit goods/drugs, invasions of privacy, disclosures of others' private information, workplace deviance, and failing to honor

Table 4
Coding of how frequently a given moral violation appeared in the 122 Stories from Test Set 2. Stories could be coded as having multiple violation types. Coding scheme was adopted from Powell & Horne (2017).

Moral Violations	Percentage of Stories Featuring Behavior
Lying	56.56%
Other	20.49%
Shoplifting	10.66%
Robbery	9.84%
Assault	4.92%
Embezzlement	3.28%
Slander	2.46%
Bullying	1.64%
Trespassing	1.64%
Adultery	0.82%
Vandalism	0.82%
Murder; Sexual Assault; Kidnapping; Arson; Incest; Drunk-Driving; Car Theft;	0.00%

a promise. The text of all stories in Test Set 2 can be found on the OSF (https://osf.io/3xt48/?view_only=a816692bb4fc4f5b87cb8e6373dfe4e7).

8. Study 3

Study 3 was embedded within the collection of Test Set B. Study 3 served as a near direct replication of Studies 1–2 with a nationally representative sample. Study 3 was preregistered (https://osf.io/bgv85).

8.1. Methods

8.1.1. Participants and design

Participants in Study 3 ($N = 230$) were the same sample as Test Set B. The only additional note is that all 230 participants were included in the analyses for Study 3. The collection of participants' stories in Test Set B was not related to their judgments of the stories from Test Set A, meaning that in Study 3 we are analyzing the 24 participants who did not consent to having their stories shared, along with the 84 participants whose stories were not chosen as stimuli due to lacking sufficient coherence and/or specificity.

8.1.2. Procedure

After providing their stories, corresponding judgments, and responses to the trait items (see Test Set B) participants shifted from being actors to being observers. The procedure was nearly identical to that of Study 2 where participants were shown a story and asked to rate the actor's motives on the six motive measures actors responded to from Test Set A. The procedure deviated from that of Study 2 in only two ways. First, participants read six instead of five stories, chosen randomly. Six stories were chosen due to survey length constraints. Study 3 was 20 min in length compared to Study 2's 15-min length, however that additional time was largely dedicated to collecting participants' stories as part of Test Set B's procedures. Our contract to collect data through Qualtrics Panels stipulated the survey take no more than 20 min to complete, and piloting suggested that providing participants with six randomly chosen stories would reach that time limit after Test Set B's procedures were included. Second, the prompt observers received was modified from that of Study 2. In Study 2 observers were effectively given the prompt that the actors themselves received, meaning that observers knew that the actors were writing about a time when they did something "bad" for a "good reason." This was originally done for the sake of transparency for observers, but it is possible giving observers this information biased their attributions of the actors' motives. As such, in Study 3 participants instead received the following information immediately prior to reading the stories: "Instructions: Below is a story written by a participant in a previous study about things they had done in real life. Please read the story and answer the questions regarding your thoughts."

8.1.3. Analyses

The analyses for Study 3 were meant to directly replicate and extend the analyses from Study 2. See Section1 of the Supplementary Materials for detailed modeling information. All the preregistered analyses from Study 2 which were adapted from the social accuracy model were also preregistered for Study 3: the confirmatory analyses of baseline accuracy, distinctive accuracy, and insight for actors and observers, and exploratory analyses of trait moderators of distinctive accuracy (see the Analysis section of Study 2 for details). Study 3 had one additional trait that was not measured in Study 2: Propensity to Engage in Unethical Workplace Behaviors (Chen & Tang, 2006). This was included to examine as a moderator of observer and meta-accuracy in addition to Machiavellianism. The previously observed negative relationship between Machiavellianism and judgment accuracy in Study 2 was likely explained by one of two possibilities: Machiavellians are also lower on empathy, which we found to be positively associated with judgment accuracy, or Machiavellians are more likely to engage in immoral behaviors and that propensity is associated with a lower ability to judge others' moral motives. We included the Propensity to Engage in Unethical Workplace Behaviors Scale to more precisely test this latter possibility. Lastly, unlike Study 2, Study 3 did not measure any other judgments of the stories (e.g. the immorality of the behavior); rather, it measured only the six motive judgments from Test Set A. Complete regression tables, including variance-covariance matrices, for all the models reported below can be found in Section 3 of the Supplemental Materials.

8.1.4. Sensitivity analyses

We conducted Monte Carlo simulations, using the *simr* R package (Green & MacLeod, 2016), of the models used to test our key hypotheses to examine the sensitivity with which they would be able to observe effect sizes at and below the observed estimate. We examined the sensitivity of the distinctive meta-accuracy and distinctive observer accuracy models. Sample size for the simulations was $N_{actors} = 13$ and $N_{observers} = 230$ (8277 actor-observer judgments). We estimated the statistical power of the observed distinctive meta-accuracy effect size ($b = 0.17$, see below), and the sensitivity to detect smaller estimates (0.12 and 0.06), based on 500 simulations for each estimate, with alpha = 0.05. The observed estimate of $b = 0.17$ was powered at 98%, 95% CI = [0.97, 0.99], while $b = 0.12$ was observable with 77% power, 95% CI = [0.73, 0.81], and $b = 0.06$ observable with 29% power, 95% CI = [0.25, 0.33]. We estimated the statistical power of the observed distinctive observer-accuracy effect size ($b = 0.26$, see below), and the sensitivity to detect smaller estimates (0.20 and 0.15), based on 500 simulations for each estimate. The observed estimate of $b = 0.26$ was powered at 86%, 95% CI = [0.83, 0.89], while $b = 0.20$ was observable with 63% power, 95% CI = [0.58, 0.67], and $b = 0.15$ was observable with 40% power, 95% CI = [0.36, 0.45].

8.2. Results

8.2.1. Meta-perception accuracy

Our results replicated the findings from Studies 1 and 2, where actors displayed both baseline meta-accuracy, $b = 0.32$, 95% CI = [0.22, 0.42], $B = 0.36$, $t(12.04) = 6.94$, $p < 0.001$, distinctive meta-accuracy, $b = 0.17$, 95% CI = [0.08, 0.25], $B = 0.20$, $t(12.00) = 4.11$, $p = 0.001$, and normative accuracy, $b = 0.86$, 95% CI = [0.36, 1.37], $B = 0.40$, $t(12.01) = 3.71$, $p = 0.003$. We found no statistically significant evidence for meta-insight, $b = 0.02$, 95% CI = [−0.01, 0.06], $B = 0.02$, $t(12.04) = 1.63$, $p = 0.129$, with meta-perceptions instead displaying a transparency bias, $b = 0.80$, 95% CI = [0.59, 1.00], $B = 0.84$, $t(11.94) = 8.52$, $p < 0.001$.

8.2.2. *Observer accuracy*

Replicating the findings from Studies 1 and 2 the observer accuracy models found that observers were able to accurately assess the self-reported motives of the actors in the stories they read across baseline accuracy, $b = 0.42$, 95% CI = [0.27, 0.58], $B = 0.36$, $t(12.01) = 6.00$, $p < 0.001$, distinctive accuracy, $b = 0.26$, 95% CI = [0.09, 0.44], $B = 0.23$, $t(11.92) = 3.23$, $p = 0.007$, and normative accuracy, $b = 0.50$, 95% CI = [0.29, 0.71], $B = 0.22$, $t(13.17) = 5.10$, $p < 0.001$. We found no evidence for actor insight, $b = 0.06$, 95% CI = [−0.22, 0.33], $B = 0.07$, $t(10.81) = 0.41$, $p = 0.693$.

8.2.3. *Moderators of observer accuracy*

To examine moderators of observer accuracy we separately entered each trait as an interaction with actors' true motives in predicting observers' distinctive accuracy (and as an interaction with the normative profile as a control). We found that cognitive ability, $b = 0.02$, 95% CI = [0.01, 0.03], $B = 0.05$, $t(224.45) = 3.64$, $p < 0.001$, and empathic-concern, $b = 0.04$, 95% CI = [0.01, 0.07], $B = 0.04$, $t(226.02) = 3.07$, $p = 0.002$, positively moderated distinctive accuracy, meaning observers higher on those traits were more accurate in their motive judgments. Conversely, Machiavellianism, $b = -0.05$, 95% CI = [−0.07, −0.03], $B = -0.06$, $t(228.20) = -4.46$, $p < 0.001$, and Propensity for Unethical Workplace Behavior, $b = -0.04$, 95% CI = [−0.07, −0.02], $B = -0.05$, $t(221.24) = -3.49$, $p = 0.001$, negatively moderated distinctive accuracy. This means that observers who self-reported engaging in more unethical workplace behaviors were less accurate in their judgments of the unique moral motives of other morally questionable actors. In other words, engaging in morally questionable behaviors does not give one superior insight into the motives of other morally questionable actors, in fact it seems to diminish one's capacity to understand why others engage in morally questionable behavior. Trait-perspective taking did not moderate observers' distinctive accuracy, $b = 0.03$, 95% CI = [−0.002, 0.06], $B = 0.03$, $t(232.86) = 1.86$, $p = 0.066$. Overall these results replicated the findings from Study 2.

9. Study 4

Study 4 involved taking the 122 stories collected in Test Set B and giving them to a new nationally representative sample of observers. Study 4 was designed to replicate and expand upon the findings from previous studies and was preregistered (https://osf.io/kzybe).

9.1. *Methods*

9.1.1. *Participants and design*

256 participants (Mean$_{\text{age}}$ = 49 years, 158 Women, 98 Men) were recruited to participate through Qualtrics survey panels. The study was advertised as taking 17–20 min and participants were paid a predetermined amount of credits through Qualtrics' internal credit system. A quota-matching system was utilized to guarantee the final sample would be representative of the general US population, with quotas set to census distributions along the following demographic characteristics: age, gender, ethnicity, education, and income. The goal was to collect 190 participants to detect our anticipated effects with 80% power, based on a power analysis using data from our previous studies (see preregistration). However, during the soft launch of data collection (at which point we had collected 43 responses) we found that one secondary task, the Raven's Matrices (Raven, 2000), was taking many participants so long to complete that the mean study length was 35 min. As such, we removed the Raven's Matrices and launched the study again, hence a total of 256 participants. When this decision was made, we preregistered the plan before we continued data collection (https://osf.io/3jnfk. We retained and analyzed the data from the first 43 participants, except for their responses to the Raven's Matrices. The 265 observers were then merged with the 122 actors from Test Set B, such that observers' judgments could be directly compared to actors' meta-perceptions.

9.1.2. *Procedure*

The procedure of Study 4 was very similar to Studies 2 and 3. Participants began by providing informed consent, responding to demographic questions for the quota-matching, then responded to an attention check and were prevented from continuing if they failed. Participants then read and rated nine stories (randomized) from Test Set B, with the same minimal instructions as were used in Study 3. For each story participants provided their motive attributions along the 16 motive items and the 11 judgment items measured in Test Set B. Following the nine stories, participants completed a comprehension check and were prevented from continuing if they failed. Then participants completed several trait measures: the perspective taking ($\alpha = 0.75$) and empathic-concern ($\alpha = 0.77$) subscales of the Interpersonal Reactivity Index (Davis, 1983), trait Machiavellianism ($\alpha = 0.88$) (Dahling et al., 2009), and Propensity to Engage in Unethical Workplace Behavior ($\alpha = 0.94$) (Chen & Tang, 2006). Participants then provided political orientation, were given the opportunity to comment on their experience, and the study ended.

9.1.3. *Analyses*

The analyses for Study 4 were meant to directly replicate and extend the analyses from Study 3. See Section1 of the Supplementary Materials for detailed modeling information. All the preregistered analyses from Study 3 which were adapted from the social accuracy model (Biesanz, 2010) were also preregistered for Study 4: the confirmatory analyses of baseline accuracy, distinctive accuracy, and insight for actors and observers, exploratory analyses of how accuracy and bias differed by motive-type, and exploratory analyses of trait moderators of accuracy. All analyses used linear mixed-effects modeling, per the social accuracy model. Complete regression tables, including variance-covariance matrices, for all the models reported below can be found in Section 3 of the Supplemental Materials.

9.1.4. *Sensitivity analyses*

We conducted Monte Carlo simulations, using the *simr* R package (Green & MacLeod, 2016), of the models used to test our key hypotheses to examine the sensitivity with which they would be able to observe effect sizes at and below the observed estimate. We examined the sensitivity of the distinctive meta-accuracy and distinctive observer accuracy models. Sample size for the simulations was $N_{actors} = 122$ and $N_{observers} = 256$ (62,199 actor-observer judgments). We estimated the statistical power of the observed distinctive meta-accuracy effect size ($b = 0.17$, see below), and the sensitivity to detect smaller estimates (0.12 and 0.06), based on 500 simulations for each estimate, with alpha = 0.05. The observed estimate of $b = 0.17$ was powered at 100%, 95% CI = [0.99, 1.00], while $b = 0.12$ was observable with 100% power, 95% CI = [0.99, 1.00], and $b = 0.06$ observable with 94% power, 95% CI = [0.90, 0.95]. We estimated the statistical power of the observed distinctive observer-accuracy effect size ($b = 0.19$, see below), and the sensitivity to detect smaller estimates (0.12 and 0.06), based on 500 simulations for each estimate. The observed estimate of $b = 0.19$ was powered at 100%, 95% CI = [0.99, 1.00], while $b = 0.12$ was observable with 100% power, 95% CI = [0.99, 1.00], and $b = 0.06$ was observable with 82% power, 95% CI = [0.78, 0.85].

9.2. *Results*

9.2.1. *Meta-perception accuracy & bias*

Our results largely replicated the findings from Studies 1–3, where actors displayed meta-accuracy, $b = 0.25$, 95% CI = [0.21, 0.28], $B = 0.26$, $t(121.27) = 14.71$, $p < 0.001$, distinctive meta-accuracy, $b = 0.17$, 95% CI = [0.13, 0.20], $B = 0.17$, $t(121.71) = 10.04$, $p < 0.001$, and normative accuracy, $b = 0.79$, 95% CI = [0.66, 0.93], $B = 0.29$, $t(121.14) = 11.53$, $p < 0.001$. Unlike Study 3, here we observed small levels of meta-insight, $b = 0.09$, 95% CI = [0.07, 0.11], $B = 0.09$, $t(120.83) = 8.25$, $p < 0.001$, suggesting that actors do have some knowledge as to how they will be misperceived by observers. Replicating Study 3, actors displayed a transparency bias, $b = 0.61$, 95% CI = [0.57, 0.66], $B = 0.63$, $t(118.66) = 27.91$, $p < 0.001$.

To examine systematic biases in meta-perceptions we variable-centered the data and interacted motive-item, as a categorical variable, with meta-accuracy in the base meta-accuracy model and computed marginal means. Variable-centering the data meant that estimates which significantly differed from zero could be interpreted as systematic directional bias within a specific perception. Results can be seen in Fig. 1.

Fig. 1 provides strong evidence for a positivity bias in moral meta-perception. Nearly all the attributions which actors systematically overestimated (estimates to the right side of the dotted line) were positively-valenced attributions, suggesting actors believed observers will perceive them and their motives more positively than observers actually did. Conversely, almost all the attributions which actors systematically underestimated were negatively-valenced, including perceptions of how wrong, unethical, immoral, and harmful actors' behavior was perceived by observers.

9.2.2. *Moderators of meta accuracy*

To examine moderators of meta-accuracy we separately entered each trait as an interaction with observers' true perceptions in predicting actors' distinctive meta-accuracy (and as an interaction with the normative profile as a control). We found that meta-cognitive skill positively moderated distinctive meta-accuracy, $b = 0.36$, 95% CI = [0.08, 0.63], $B = 0.04$, $t(119.91) = 2.59$, $p = 0.011$. Meta-cognitive skill was measured as actors' distinctive accuracy random slope estimates as *observers* in Study 3, meaning that participants who exhibited greater distinctive observer-accuracy were also more accurate meta-perceivers. Other traits did not moderate distinctive meta-accuracy, namely perspective-taking, $b = -0.02$, 95% CI = [−0.05, 0.01], $B = -0.02$, $t(126.34) = -1.16$, $p = 0.249$, empathic-concern, $b = 0.02$, 95% CI = [−0.01, 0.05], $B = 0.02$, $t(120.24) = 1.20$, $p = 0.222$, cognitive ability, $b = 0.00$, 95% CI = [−0.01, 0.02], $B = 0.01$, $t(120.06) = 0.72$, $p = 0.470$, propensity for workplace unethical behavior, $b = -0.01$, 95% CI = [−0.04, 0.02], $B = -0.01$, $t(118.79) = -0.47$, $p = 0.636$, and Machiavellianism, $b = -0.02$, 95% CI = [−0.05, 0.01], $B = -0.03$, $t(120.07) = -1.62$, $p = 0.107$.

To examine whether the severity of the morally questionable behavior moderated meta-accuracy, we performed two additional moderation analyses. We averaged ratings of immorality (the "immoral," "wrong" and "unethical" items) for each story, separately for observers' perceptions and actors' self-perceptions, and centered each distribution on the scale midpoint (4) as the scale was bipolar (e.g., "Very Moral" to "Very Immoral"). We then interacted each separately with distinctive meta-accuracy and normative accuracy. We found that observers' perceived immorality positively moderated distinctive meta-accuracy, $b = 0.04$, 95% CI = [0.01, 0.07], $B = 0.05$, $t(114.75) = 2.57$, $p = 0.012$, as did actors' self perceptions of immorality, $b = 0.03$, 95% CI = [0.01, 0.07], $B = 0.05$, $t(120.12) = 3.00$, $p = 0.003$, meaning that actors whose behaviors were seen as immoral and who saw their own behavior as immoral had higher distinctive meta-accuracy. Because of the covariance between self- and other-perception of immorality, we lastly entered both as moderators into a single model. There were no concerns over multicollinearity (all VIFs <= 1.73), and actors' self-perceived immorality continued to positively moderate distinctive meta-accuracy, $b = 0.02$, 95% CI = [0.002, 0.04], $B = 0.03$, $t(119.50) = 2.15$, $p = 0.033$, whereas observers' perceived immorality did not, $b = 0.02$, 95% CI = [−0.01, 0.06], $B = 0.03$, $t(118.78) = 1.43$, $p = 0.155$.

9.2.3. *Observer accuracy*

Replicating the findings from Studies 1–3, the observer accuracy models found that observers were able to accurately assess the self-reported motives of the actor in the stories they read in the baseline, $b = 0.26$, 95% CI = [0.22, 0.30], $B = 0.24$, $t(157.88) = 12.43$, $p < 0.001$, distinctive accuracy, $b = 0.19$, 95% CI = [0.15, 0.23], $B = 0.18$, $t(145.11) = 9.27$, $p < 0.001$, and insight models, $b = 0.14$, 95% CI = [0.10, 0.18], $B = 0.13$, $t(136.62) = 6.58$, $p < 0.001$. We also found that observers' judgments related to both the normative profile, $b = 0.45$,

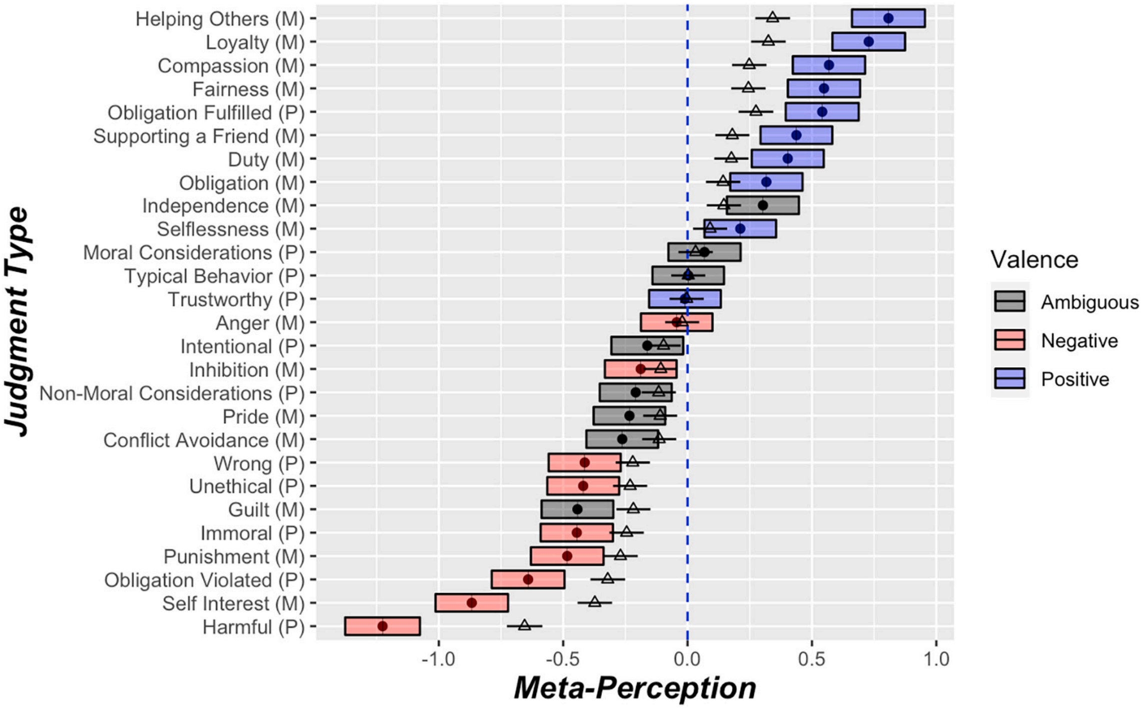

Fig. 1. Marginal mean estimates of meta-perceptions. Black dots with colored bars are unstandardized estimates, triangles are standardized beta estimates, all bars are 95% confidence intervals. Values were true-mean centered within judgment type, such that zero (the dotted vertical line) was interpreted as mean-level accuracy, and estimates which deviated from zero were interpreted as directional bias (greater than zero represented overestimation, and vice versa). Judgments labeled "M" are motive attributions, and "P" perceptions of the actor/behavior. Valence labels were generated by the researchers.

95% CI = [0.36, 0.54], $B = 0.16$, $t(175.75) = 9.77$, $p < 0.001$, and the opaqueness of actors' motives, $b = 0.18$, 95% CI = [0.14, 0.23], $B = 0.17$, $t(127.35) = 8.14$, $p < 0.001$.

9.2.4. Moderators of observer accuracy

To examine moderators of observer-accuracy we separately entered each trait as an interaction with actors' true motives in predicting observers' attributions in the distinctive actor accuracy model (and as an interaction with the normative profile as a control). We found that trait perspective-taking, $b = 0.03$, 95% CI = [0.02, 0.04], $B = 0.03$, $t(253.77) = 4.28$, $p < 0.001$, and empathic-concern, $b = 0.04$, 95% CI = [0.03, 0.06], $B = 0.04$, $t(247.25) = 6.62$, $p < 0.001$, positively moderated distinctive accuracy, whereas Machiavellianism, $b = -0.03$, 95% CI = [−0.04, −0.02], $B = -0.04$, $t(250.76) = -5.07$, $p < 0.001$, and Propensity for Workplace Unethical Behavior, $b = -0.05$, 95% CI = [−0.06, −0.03], $B = -0.05$, $t(249.70) = -7.14$, $p < 0.001$, negatively moderated distinctive accuracy, broadly replicating the findings from Study 3.

10. General discussion

Studies 1–4 found that morally questionable actors understood how their behaviors would be uniquely perceived by others, here defined as distinctive meta-accuracy. Study 4 found that distinctive meta-accuracy was greater for those higher in cognitive ability and those who were more accurate in judging the moral motives of others, providing evidence for general moral meta-cognitive skill across tasks. Studies 1, 2 and 4 found evidence that morally questionable actors had knowledge of how they would be *misperceived* by observers, here defined as insight (Carlson, 2016). Study 3 found no statistically significant insight slope, although the coefficient was positive. Studies 1–4 also found that morally questionable actors had knowledge of the motives possessed by other morally questionable actors, here defined as normative accuracy (Furr, 2008). These results collectively suggest that individuals who have engaged in morally questionable behavior have generalized meta-accuracy (Kenny & DePaulo, 1993) of how third-parties will react to their moral violations.

Studies 1–4 also found similar evidence for accuracy in observers' judgments of the unique motives of morally questionable actors, suggesting that individuals are able to successfully perspective-take with those who have committed moral violations. Observers higher in cognitive ability (Studies 2–3) and empathic concern (Studies 2–4) were consistently more accurate in these judgments, while observers higher in Machiavellianism (Studies 2–4) and the propensity to engage in unethical workplace behaviors (Studies 3–4) were consistently less accurate. This latter result suggests that more frequently engaging in morally questionable behavior does not grant one insight into the moral minds of others, and in fact is associated with less ability to understand the motives behind others' morally questionable behavior.

Despite strong evidence for generalized meta-accuracy (and observer accuracy) across studies, actors' accuracy in judging how they would be perceived was accompanied by two judgment biases. Studies 1–4 found evidence for a transparency bias among morally questionable actors (Gilovich et al., 1998), meaning that actors *overestimated* how accurately observers would perceive their self-reported moral motives. Similarly, in Study 4 an examination of actors' meta-perception point estimates found evidence for a positivity bias. Actors systematically overestimate the positive attributions, and underestimate the negative attributions, made of them and their motives. In fact, the single meta-perception found to be the most inaccurate in its average point estimate was the meta-perception of harm caused, which was significantly underestimated. In short, meta-perception across motive judgments was accurate, but point estimate accuracy in judging specific motives displayed strong valence biases.

We also observed consistent moderators of observer-accuracy, examined in Studies 2–4, and moderators of meta-accuracy, examined in Study 4. Among observers, cognitive ability and empathic-concern were consistent positive predictors of distinctive-accuracy, whereas Machiavellianism and Propensity for Workplace Unethical Behavior were consistent negative predictors of distinctive-accuracy. If we assume that those high on Machiavellianism and Propensity for Workplace Unethical Behavior have engaged in more morally questionable actions than individuals low on those traits, then our findings suggest that such experiences actually attenuate the ability to understand others' moral motives, rather than providing greater insight into what motivates others' morally questionable actions.

The only individual-difference measure found to moderate distinctive meta-accuracy was actors' accuracy as observers judging the motives of others, which we interpret as evidence for general meta-cognitive skill, although this could also be partially explained by systematic differences in how individuals explain their own moral behaviors (Capraro, Vanzo, & Cabrales, 2022). Given the anticipated variance in the severity of morally questionable acts, we also examined self- and other judgments of immorality and found that actors who rated their own behavior as immoral had greater distinctive meta-accuracy than actors who self-rated as moral. There are several theoretical explanations for this relationship. One is that greater willingness to label one's own actions as immoral correlates with traits which themselves may predict more accurate meta-perception, such as trait empathy or humility. Another possibility is that those who see their own actions as immoral have greater levels of guilt, which led them to provide greater explanation for their own behavior in their stories, allowing greater meta-accuracy to arise from greater shared information between actors and observers. Similarly, it may be that actors who self-rated as immoral wrote about behavior which, despite the prompt to participants, contain few “good” reasons, making their motives less ambiguous overall.

This work makes three parallel contributions. First, we tested for different components of accuracy across a host of moral judgments, providing detailed and generalizable accounts of how morally questionable actors believed they would be perceived, how they were actually perceived, and the role of individual differences in meta-accuracy. Second, by focusing on discrete moral motives (e.g. anger, loyalty, conflict avoidance, etc.) as the basis of observer and meta-judgments, we contribute to a nascent body of research examining how (im)moral actors describe their own behavior and how motive attributions relate to perceptions of wrongdoing (Ames & Fiske, 2013; Rai & Fiske, 2011; Young & Saxe, 2011). To our knowledge our findings are the first to ask morally questionable actors and observers to rate behaviors on a large and theoretically diverse set of moral motives, rather than just intentionality or a narrow set of motivational attributions. Our choice of motive items required participants to make assumptions about the breadth and totality of relevant motives for the average morally questionable actor and observer. Further theorizing and empirical work is needed to better understand both the motives morally questionable actors ascribe to themselves, and the motives observers naturalistically ascribe to other's morally questionable behaviors.

Third, this work introduced a novel methodology for capturing naturalistic moral behavior and judgment. We asked individuals to provide written accounts of past morally questionable behavior, then to rate their own motives and the motives they believe others would attribute to them (meta-perception). We then had a separate sample of observers read these accounts, provide their motive attributions, and compared those directly to the morally questionable actors' meta-perceptions, allowing for a direct test of meta-accuracy. This approach produced a large and diverse set of naturalistic instances of moral behavior, and allowed for more generalizable inferences regarding moral judgment (Yarkoni, 2020) compared to past work which has typically relied on highly stylized and fictional moral scenarios (e.g., sacrificial dilemmas) that have been increasingly criticized for lacking generalizability (Bauman, McGraw, Bartels, & Warren, 2014; Graham, 2014).

Our findings also highlight the benefits of heeding calls for componential rather than univariate analyses of judgment accuracy (Biesanz, 2010; Lees & Cikara, 2021; West & Kenny, 2011; Wood & Furr, 2016). Simple methods of operationalizing accuracy which are still common in psychology, such as bivariate correlations and *t*-tests, would have masked a wealth of information about the nature of moral meta-perception. The conception that judgment can be *both* accurate and biased simultaneously is born out in our findings. By embracing that theoretical and empirical framework, we uncovered a richer and more informative examination of whether morally questionable actors know how they will be perceived by others.

A theoretical assumption embedded within our componential approach is that all actors share a similar set of normative expectations regarding how they are judged, reflected in our operationalization of the normative profile. It is possible a morally questionable actor who stole something had distinct normative expectations regarding how they would be judged relative to a morally questionable actor who lied, though the lack of existing theorizing regarding how morally questionable actors expect to be perceived by others makes such a possibility uncertain. While our findings provide evidence that the paradigm-embedded normative profile is strongly associated with generalized meta-perceptions (and observer perceptions), we are unable to test for the possibility that there is systematic and qualitative variance in the normative expectations held by actors engaging in different types of behaviors.

Another theoretical assumption inherent to our paradigm is that actors describe behaviors which meet their *personal* definition of "something bad for a good reason." As such, we are reluctant to label any of the actions as lacking a "good reason" *ex post*, either categorically or based on actors' descriptions. Actors' self-perceptions reflect the ambivalence one would expect from accounts of "bad" behavior for a "good reason." Self-perceptions of immorality were asked on a bipolar scale (e.g., 1–7 scale where 1 is labeled "Very Moral" and 7 "Very Immoral"), and the median self-perception is exactly 4.0 (Mean = 3.9, SD = 1.7), meaning actors were nearly evenly split as to whether their actions were on average moral vs. immoral. And while "bad behaviors for bad reasons" are outside of the scope of the current research, our findings suggest that meta-perceptions for such behaviors would not be categorically distinct from those of morally questionable behaviors. First, participants who are theoretically more likely to engage in immoral behaviors (Machiavellians and Workplace Deviants) were not more or less accurate in their meta-perceptions and were slightly less accurate in judging the motives of others. Moreover, in Study 4, actors who rated themselves as more immoral were slightly more accurate in their distinctive meta-perceptions than actors who rated their behavior as moral overall. These findings suggest that while propensity and severity of morally questionable and immoral actions may play a role in meta- and observer-accuracy, they do not categorically shift either.

In addition to this work's contributions, several limitations of these studies point to fruitful avenues of future research. First, Test Set A and Studies 1–2 used convenience samples, meaning that the "true values" derived therein should be interpreted with some cautions regarding their representativeness. And while Test Set B was collected with a nationally representative sample, the necessary culling of the stories down to a smaller set likely meant that the actor set used in Study 4 was not perfectly representative.

Second, the finding in Study 4 of a strong positivity bias was an exploratory hypothesis. The analysis was preregistered, but the positivity bias was not predicted *ex ante*. We argue that the positivity bias is empirically unambiguous, and while positive self-evaluation exists in moral (Tappin & McKay, 2017) and non-moral domains (Alicke, 1985), positive *self*-evaluation does not necessarily entail positive *meta*-perception. Indeed, some work suggests people are unaware of the positive impression others hold of them (Gallrein et al., 2016). Moreover, the positivity bias here should be interpreted in context: the paradigm regarded morally questionable behaviors, actors on average rated their own behaviors as immoral, actors' meta-perceptions were more negative than their self-perceptions, and actors had perfect knowledge of what information observers would have. Yet despite this, actors *still* exhibited a striking positivity bias in meta-perception. As such, this positivity bias should not be assumed an extension of a mere positivity bias in self-evaluation.

Third, given that the paradigm is novel it is difficult to know how participants' accounts of their behavior, and self/meta-perceptions, might differ from other possible paradigms. For example, writing about one's wrongdoings versus conveying them face-to-face to another person might enable participants to be more open about their motives and behavior, and therefore our paradigm may be partially inducing the observed transparency bias. Face-to-face recountings of morally questionable behavior would also constitute dyadic meta-accuracy, rather than the generalized meta-accuracy we examine. Future work should strongly consider alternative paradigms to better assess how the format and context might affect participants' accounts and perceptions, and how moral meta-perception may differ at the dyadic vs. generalized level. This work also relied on assumptions regarding the comprehensiveness of the list of motives (namely in Test Set B). Future work should consider more grounded approaches to developing lists of relevant moral motives, or consider adding motives which were not present in Test Set B, for example purity motives (Chakroff & Young, 2015; Graham et al., 2013).

It is worth considering the paradigmatic similarities between our method and the sacrificial dilemmas commonly used in moral psychology. Similar to sacrificial dilemmas, our stories represent moral "gray" areas where actors are often faced with competing and irreconcilable moral motives. Yet while sacrificial dilemmas prioritized internal validity, and been criticized for a lack of realism (Bauman et al., 2014; Bostyn, Sevenhant, & Roets, 2018; Hester & Gray, 2020), our paradigm prioritizes generalizability at the expense of experimental control, which can make direct comparisons between observers' attributions across our stories and past work on sacrificial dilemmas difficult. Moreover, sacrificial dilemmas are designed to answer research questions embedded within a dual-process framework of moral cognition, whereas our open-ended paradigm of generating stories is designed to answer more proximate questions of judgment accuracy which, while broadly generalizable, are less embedded within established theories of moral cognition. Nonetheless, one pattern of accuracy that is consistent with a dual-process account of moral judgment is the directional effects of pro- and antisocial traits. We consistently found that empathic-concern predicted greater levels of observer accuracy, whereas Machiavellianism and Propensity for Workplace Unethical Behavior were consistent predictors of lower observer accuracy. Similarly, pro-social traits, such as empathic concern, are associated with greater deontological decision-making (Cameron, Conway, & Scheffer, 2022; Nasello, Dardenne, Blavier, & Triffaux, 2021; Reynolds & Conway, 2018), while antisocial traits, such as psychopathy and generalized distrust, are associated with lower deontological decision-making (Conway, Weiss, Burgmer, & Mussweiler, 2018; Luke & Gawronski, 2021), although these relationships are complex and can reverse (Fleischmann, Lammers, Conway, & Galinsky, 2019). Machiavellianism specifically can also be associated with utilitarian judgments, but only when such judgments are self-interested (Zamora, Ungson, & Seidman, 2022). As such, dual-process accounts of moral judgment may cohere with the pattern of observer accuracy moderators we find, although our studies were not designed to answer such research questions.

Lastly, while not the focus of this research, there is good reason to believe our results do not generalize to the *victims* of morally questionable behaviors. Our observers were neutral third-parties, and actors were asked for meta-perception of "the average person" (i.e., generalized meta-accuracy) not the victim or anyone involved in their stories. Our results should not be interpreted as evidence that morally questionable actors have knowledge of how the victims of their behavior perceive them.

By integrating theory and methods across moral psychology and social perception we demonstrate a robust pattern of moral meta-perception accuracy, advance our understanding of moral cognition, and provide new methods for moral psychologists seeking to understand the accuracy of (meta)moral judgments.

Funding

The research was generously funded by Harvard Business School's Doctoral Research Fund. Funds were utilized to compensate survey participants.

Ethics

This research was approved by Harvard University's Institutional Review Board. All participants provided informed consent to participate, and no deception was utilized in this study.

Contributions

J.L. initially proposed the research question, and all authors contributed equally to the study designs and drafting of the manuscript. J.L. prepared the online surveys, facilitated the preregistrations, data collection, analyses, and public posting of materials.

Acknowledgments

We thank Harvard Business School's Organizational Behavior Lab and Boston College's Morality Lab for their helpful feedback on study design, Steven Worthington and Ista Zahn at Harvard's IQSS for statistical guidance, and Haley Todd for coding the stories.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104371.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49*(6), 1621–1630. https://doi.org/10.1037/0022-3514.49.6.1621

Ames, D., & Fiske, S. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24*(9).

Ames, D., & Fiske, S. (2015). Perceived intent motivates people to magnify observed harms. *Proceedings of the National Academy of Sciences, 112*(12), 3599–3605. https://doi.org/10.1073/pnas.1501592112

Barasch, A., Levine, E. E., Berman, J. Z., & Small, D. A. (2014). Selfish or selfless? On the signal value of emotion in altruistic behavior. *Journal of Personality and Social Psychology, 107*(3), 393–413. https://doi.org/10.1037/a0037207

Baron, J., Ritov, I., & Greene, J. D. (2011). The duty to support nationalistic policies. *Journal of Behavioral Decision Making, 11*.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barranti, M., Carlson, E. N., & Furr, R. M. (2016). Disagreement about moral character is linked to interpersonal costs. *Social Psychological and Personality Science, 00*(0), 1–12. https://doi.org/10.1177/1948550616662127

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*(1), 154–161. https://doi.org/10.1016/j.cognition.2011.05.010

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology, 68*(6), 1042–1054.

Bauman, C., McGraw, A., Bartels, D., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*(9), 536–554. https://doi.org/10.1111/spc3.12131

Biesanz, J. C. (2010). The social accuracy model of interpersonal perception: Assessing individual differences in perceptive and expressive accuracy. *Multivariate Behavioral Research, 45*(5), 853–885. https://doi.org/10.1080/00273171.2010.519262

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084–1093.

Bursztyn, L., Gonzalez, A. L., & Yanagizawa-Drott, D. (2018). *Misperceived social norms: Female labor force participation in Saudi Arabia*. National Bureau of Economic Research Working Paper Series.

Cameron, C. D., Conway, P., & Scheffer, J. A. (2022). Empathy regulation, prosociality, and moral judgment. *Current Opinion in Psychology, 44*, 188–195. https://doi.org/10.1016/j.copsyc.2021.09.011

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: No evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review, 19*(4), 371–394. https://doi.org/10.1177/1088868314566683

Capraro, V., Sippel, J., Zhao, B., Hornischer, L., Savary, M., Terzopoulou, Z., … Griffioen, S. F. (2018). People making deontological judgments in the Trapdoor dilemma are perceived to be more prosocial in economic games than they actually are. *PLoS One, 13*(10), Article e0205066. https://doi.org/10.1371/journal.pone.0205066

Capraro, V., Vanzo, A., & Cabrales, A. (2022). Playing with words: Do people exploit loaded language to affect others' decisions for their own benefit? *Judgment and Decision making, 47*(1). https://doi.org/10.31234/osf.io/yswxe

Carlson, E. N. (2016). Meta-accuracy and relationship quality: Weighing the costs and benefits of knowing what people really think about you. *Journal of Personality and Social Psychology, 111*(2), 250–264. https://doi.org/10.1037/pspp0000107

Carlson, E. N., & Furr, R. M. (2009). Evidence of differential meta-accuracy: People understand the different impressions they make. *Psychological Science, 20*(8), 1033–1039.

Carlson, E. N., Vazire, S., & Oltmanns, T. F. (2011). You probably think this paper's about you: Narcissists' perceptions of their personality and reputation. *Journal of Personality and Social Psychology, 101*(1), 185–201. https://doi.org/10.1037/a0023781.You

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition, 136*, 30–37. https://doi.org/10.1016/j.cognition.2014.11.034

Chen, Y.-J., & Tang, T. L.-P. (2006). Attitude toward and propensity to engage in unethical behavior: Measurement invariance across major among university students. *Journal of Business Ethics, 69*(1), 77–93. https://doi.org/10.1007/s10551-006-9069-6

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*(2), 216–235. https://doi.org/10.1037/a0031021

Conway, P., Weiss, A., Burgmer, P., & Mussweiler, T. (2018). Distrusting your moral compass: The impact of distrust mindsets on moral dilemma processing and judgments. *Social Cognition, 36*(3), 345–380. https://doi.org/10.1521/soco.2018.36.3.345

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences, 111*(48), 17320–17325. https://doi.org/10.1073/pnas.1424572112

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292. https://doi.org/10.1177/1088868313495594

Dahling, J. J., Whitaker, B. G., & Levy, P. E. (2009). The development and validation of a new machiavellianism scale. *Journal of Management, 35*(2), 219–257. https://doi.org/10.1177/0149206308318618

Davis, M. H. (1983). A mulitdimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology, 44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Fleischmann, A., Lammers, J., Conway, P., & Galinsky, A. D. (2019). Paradoxical effects of power on moral thinking: Why power both increases and decreases deontological and utilitarian moral decisions. *Social Psychological and Personality Science, 10*(1), 110–120. https://doi.org/10.1177/1948550617744022

Funder, D. C. (1995). On the accuracy of personality judgment:A realistic approach. *Psychological Review, 102*(4), 652–670.

Furr, R. M. (2008). A framework for profile similarity: Integrating similarity, normativeness, and distinctiveness. *Journal of Personality, 76*(5), 1267–1316. https://doi.org/10.1111/j.1467-6494.2008.00521.x

Furr, R. M., & Funder, D. C. (2004). Situational similarity and behavioral consistency: Subjective, objective, variable-centered, and person-centered approaches. *Journal of Research in Personality, 38*(5), 421–447. https://doi.org/10.1016/j.jrp.2003.10.001

Gallrein, A.-M. B., Weßels, N. M., Carlson, E. N., & Leising, D. (2016). I still cannot see it – a replication of blind spots in self-perception. *Journal of Research in Personality, 60*, 1–7. https://doi.org/10.1016/j.jrp.2015.10.002

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition, 177*, 122–141. https://doi.org/10.1016/j.cognition.2018.03.019

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75*(2), 332–346. https://doi.org/10.1037/0022-3514.75.2.332

Graham, J. (2014). Morality beyond the lab. *Science, 345*(6202), 1242. https://doi.org/10.1126/science.1259500

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology, 47*, 55–130.

Green, P., & MacLeod, C. J. (2016). simr: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods, 25*(3), 365–379. https://doi.org/10.1037/met0000239

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. https://doi.org/10.1177/0146167214554957

Hester, N., & Gray, K. (2020). The moral psychology of raceless, genderless strangers. *Perspectives on Psychological Science, 15*(2), 216–230. https://doi.org/10.1177/1745691619885840

Janoff-Bulman, R., Sheikh, S., & Baldacci, K. G. (2008). Mapping moral motives: Approach, avoidance, and political orientation. *Journal of Experimental Social Psychology, 44*(4), 1091–1099. https://doi.org/10.1016/j.jesp.2007.11.003

Jordan, J. J., & Rand, D. G. (2020). Signaling when nobody is watching: A reputation heuristics account of outrage and punishment in one-shot anonymous interaction. *Journal of Personality and Social Psychology, 118*(1), 57–88. https://doi.org/10.1037/pspi0000186

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review, 8*(3), 265–280. https://doi.org/10.1207/s15327957pspr0803_3

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lees, J., & Cikara, M. (2020). Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour, 4*(3), 279–286. https://doi.org/10.1038/s41562-019-0766-4

Lees, J., & Cikara, M. (2021). Understanding and combating misperceived polarization. *Philosophical Transactions of the Royal Society B, 376*(1822), 20200143. https://doi.org/10.1098/rstb.2020.0143

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Luke, D. M., & Gawronski, B. (2021). Psychopathy and moral dilemma judgments: A CNI model analysis of personal and perceived societal standard. *Social Cognition, 39*(1), 41–58. https://doi.org/10.1521/soco.2021.39.1.41

Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B., & Bruneau, E. (2020). Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences, 117*(26), 14864–14872. https://doi.org/10.1073/pnas.2001263117

Nasello, J. A., Dardenne, B., Blavier, A., & Triffaux, J.-M. (2021). Does empathy predict decision-making in everyday trolley-like problems? *Current Psychology*. https://doi.org/10.1007/s12144-021-01566-1

Powell, D., & Horne, Z. (2017). Moral severity is represented as a domain-general magnitude. *Experimental Psychology, 64*(2), 142–147. https://doi.org/10.1027/1618-3169/a000354

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology, 64*(2), 243–256.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118*(1), 57–75. https://doi.org/10.1037/a0021867

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41*(1), 1–48. https://doi.org/10.1006/cogp.1999.0735

Reynolds, C. J., & Conway, P. (2018). Not just bad actions: Affective concern for bad outcomes contributes to moral condemnation of harm in moral dilemmas. *Emotion, 18*(7), 1009–1023. https://doi.org/10.1037/emo0000413

Rom, S. C., & Conway, P. (2018). The strategic moral self: Self-presentation shapes moral dilemma judgments. *Journal of Experimental Social Psychology, 74*, 24–37. https://doi.org/10.1016/j.jesp.2017.08.003

Rom, S. C., Weiss, A., & Conway, P. (2017). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58. https://doi.org/10.1016/j.jesp.2016.09.007

Ruggeri, K., Većkalov, B., Bojanić, L., Andersen, T. L., Ashcroft-Jones, S., Ayacaxli, N., … Folke, T. (2021). The general fault in our fault lines. *Nature Human Behaviour, 5*(10), 1369–1380. https://doi.org/10.1038/s41562-021-01092-x

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 1–39*. https://doi.org/10.1177/1088868317698288

Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods, 21*(2), 206–221. https://doi.org/10.1037/met0000062

Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*(1), 345–372. https://doi.org/10.1146/annurev.psych.56.091103.070145

Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological and Personality Science, 8*(6), 623–631. https://doi.org/10.1177/1948550616673878

Teper, R., Zhong, C.-B., & Inzlicht, M. (2015). How emotions shape moral behavior: Some answers (and questions) for the field of moral psychology. *Social and Personality Psychology Compass, 9*(1), 1–14. https://doi.org/10.1111/spc3.12154

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision making, 4*(6), 13.

Vonasch, A. J., & Sjåstad, H. (2020). Future-orientation (as trait and state) promotes reputation-protective choice in moral dilemmas. *Social Psychological and Personality Science, 194855061989925*. https://doi.org/10.1177/1948550619899257

Waytz, A., Young, L. L., & Ginges, J. (2014). Motive attribution asymmetry for love vs. hate drives intractable conflict. *Proceedings of the National Academy of Sciences, 111*(44), 15687–15692. https://doi.org/10.1073/pnas.1414146111

West, T. V., & Kenny, D. A. (2011). The truth and bias model of judgment. *Psychological Review, 118*(2), 357–378. https://doi.org/10.1037/a0022936

Wheeler, M. A., & Laham, S. M. (2016). What we talk about when we talk about morality: Deontological, consequentialist, and emotive language use in justifications across foundation-specific moral violations. *Personality and Social Psychology Bulletin, 42*(9), 1206–1216. https://doi.org/10.1177/0146167216653374

Wood, D., & Furr, R. M. (2016). The correlates of similarity estimates are often misleadingly positive: The nature and scope of the problem, and some solutions. *Personality and Social Psychology Review, 20*(2), 79–99. https://doi.org/10.1177/1088868315581119

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences, 1–37*. https://doi.org/10.1017/S0140525X20001685

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202–214. https://doi.org/10.1016/j.cognition.2011.04.005

Young, L., & Tsoi, L. (2013). When mental states matter, when they don't, and what that means for morality. *Social and Personality Psychology Compass, 7*(8), 585–604. https://doi.org/10.1111/spc3.12044

Zamora, J., Ungson, N. D., & Seidman, G. (2022). The end justifies the me: Self-interest moderates the relationship between dark triad traits and utilitarian moral decisions. *Personality and Individual Differences, 184*, Article 111134. https://doi.org/10.1016/j.paid.2021.111134