Cognition 251 (2024) 105865

Contents lists available at ScienceDirect

Cognition

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

Goal inference in moral narratives

Judy Sein Kim [a,b,*,1], Clara Colombatto [c,d,1], M.J. Crockett [a,b,*]

[a] University Center for Human Values, Princeton University, USA
[b] Department of Psychology, Princeton University, USA
[c] Department of Experimental Psychology, University College London, UK
[d] Department of Psychology, University of Waterloo, Canada

ARTICLE INFO

Keywords:
Narrative
Testimony
Reputational goals
Morality
Pragmatics

ABSTRACT

We often form beliefs about others based on narratives they tell about their own moral actions. When constructing such moral narratives, narrators balance multiple goals, such as conveying accurate information about what happened ('informational goals') and swaying audiences' impressions about their moral characters ('reputational goals'). Here, we ask to what extent audiences' detection of narrators' reputational goals guide or prevent them from making moral character judgments intended by narrators. Across two pre-registered experiments, audiences read narratives written by real narrators about their own moral actions. Each narrator was incentivized to write about the same action twice while trying to appear like a morally good or bad person (positive and negative reputational goals). Audiences detected narrators' reputational goals with high accuracy and made judgments about moral character that aligned with narrators' goals. However, audiences were more suspicious toward positive than negative reputational goals, requiring more evidence of high informational goals. These results demonstrate how audiences' inferences of reputational goals can both support and hinder narrators: accurate goal recognition increases the chance that audiences will make judgments intended by narrators, but inferred positive reputational goals can lead to doubts about accuracy. More generally, this provides a novel approach to studying how moral information about people is transmitted through naturalistic narratives.

1. Introduction

Much of our everyday communication revolves around people and their moral characters (Banerjee, Heyman, & Lee, 2020; Baumeister, Zhang, & Vohs, 2004; Bietti, Tilston, & Bangerter, 2019; Dunbar, 2004; Li & Tomasello, 2021; Pasupathi & Wainryb, 2010; Schlenker, 2003; Wiessner, 2014; Yerkovich, 1977). People share with others information about their past actions, decisions, and motivations, painting themselves as morally good or bad. These *moral narratives* are often unverifiable, describing facts that audiences could not have witnessed firsthand (e.g., imagine stories you hear when catching up with a friend, listening to a court hearing, or scrolling through social media). How do audiences make sense of other people's narratives about their moral characters? At times, they might interpret and trust a narrative just as intended by the narrator. After all, accurate comprehension and acceptance of other people's testimony allows valuable transmission of knowledge (Harris & Koenig, 2006; Harris, Koenig, Corriveau, & Jaswal, 2018; Hills, 2013; Mahr & Csibra, 2020; Sperber, 2001). At other times, however, audiences might unintentionally fail to take in a narrator's intended meaning or refuse to do so out of mistrust (Grice, 1975; Mercier, 2020; Sperber et al., 2010; Sperber & Wilson, 1986; Yus, 1999).

While most research on testimony has focused on audience comprehension and trust of simple utterances and propositions, we focus on testimony in the form of *narratives*. Narratives typically describe concrete events, actions, and mental states (e.g., Aronowitz & Lombrozo, 2020; Currie, 2010; Genette, 1980; Kim & Crockett, 2022). In contrast to a mere story which represents the facts of what happened, a narrative provides a subjective version of the story, reflecting the goals of its narrator. Narratives do not merely state the evaluations they intend for the audience (e.g., "I'm a good person"), but rather, indirectly demonstrate and provide support for them (Bietti et al., 2019; Currie, 2010; Dahlstrom, 2021; Fraser, 2021; Genette, 1980; Kim & Crockett, 2022). In the case of moral narratives, narrators often have *reputational goals* of appearing morally good or bad to others (e.g., Banerjee et al., 2020; Brown & Levinson, 1978; Schlenker, 2003; Sedikides, Hoorens, & Dufner, 2015). Narrators with different reputational goals might select

* Corresponding authors at: Department of Psychology, Princeton University, Peretsman-Scully Hall 310, Princeton, NJ 08540, USA.
E-mail addresses: judyseinkim@gmail.com (J.S. Kim), mj.crockett@princeton.edu (M.J. Crockett).
[1] These authors contributed equally.

https://doi.org/10.1016/j.cognition.2024.105865
Received 12 October 2023; Received in revised form 11 June 2024; Accepted 19 June 2024
Available online 9 August 2024

different but truthful details that they think will steer audiences toward the desired beliefs (e.g., about the narrator's own moral character). For example, a narrator with a positive reputational goal might highlight the positive outcomes of their actions while omitting negative ones (and vice versa for a negative reputational goal).

In this paper, we examine comprehension and trust of moral narratives by focusing on the role of audiences' inferences of narrators' goals. While an utterance might not explicitly state the intended evaluation, it might make an *intention* to transmit it more or less overt (e.g., Brown & Levinson, 1978; Davis, 2016; Haugh, 2009; Horn, 2010; Reboul, 2011). In the case of moral narratives, narrators might vary in how clearly they signal their desire to lead audiences to particular beliefs about moral character (i.e., make their reputational goals recognizable). For instance, a narrator who only presents positive information about themselves might make their positive reputational goal more obvious than another narrator who presents a more balanced image. Similarly, narrators might vary in how much they commit to the truth of their evaluation (e.g., signaling low vs. high confidence about their own belief that they are a good person), making their reputational goal more or less clear (cf., Bonalumi, Mahr, Marie, & Pouscoulous, 2023; Mazzarella, Reinecke, Noveck, & Mercier, 2018; Meibauer, 2014; Reboul, 2017; Vullioud, Clément, Scott-Phillips, & Mercier, 2017).

One empirical question is to what extent audiences' detection of narrators' intended meanings facilitate comprehension and acceptance (Bietti et al., 2019; Mazzarella et al., 2018; Mazzarella & Vaccargiu, 2024; Reboul, 2017; Sperber et al., 2010; Vallauri et al., 2020). On the one hand, recognition of a narrator's intentions can help align audience beliefs to what the narrator desires, avoiding miscommunication (Grice, 1957; Sperber et al., 2010; Sperber & Wilson, 1986). All utterances have multiple possible meanings, making uncertainty and noise inherent to communication. This is most evident in cases of indirect speech, where narrators do not explicitly state what they mean, but count on audiences inferring their non-literal meaning (Brown & Levinson, 1978; Searle, 1979; Yus, 1999). For example, take a narrator who tells their friend that their poem is "not terrible". They likely intended for the friend to recognize that they think the poem is quite bad but are being polite about it (Mazzarella, 2015; Yoon, Tessler, Goodman, & Frank, 2020). The friend, however, might misinterpret the narrator as wanting to communicate that they thought the poem was good (the opposite of terrible).

Moral narratives, as we have defined them, similarly allow varying interpretations because they indirectly demonstrate and provide support for particular evaluations through the selective presentation of details, rather than through direct explication. To demonstrate this point, imagine a narrator who tells the following story to his friends:

> *A few years ago, my wife accidentally cut her finger while cutting carrots, while we were on vacation… This was a big deal – you know, she's a concert pianist, she has multiple tours every year, her fingers are the most valuable thing in our household! We were staying in a cabin in the woods quite far away from the hospital. But I was so desperate to get her there quickly. We got in the car, and then I drove really fast in a frenzy, cutting red lights and ignoring crosswalks… At some point, I almost hit a kid but didn't even stop to check… And looking back, I feel really guilty about how I acted. I really could have killed someone that day, all for a finger. The choices you can make when you're really in love with someone…*

Suppose the narrator told the story because he feels genuinely guilty about the incident, believes his actions were morally wrong, and wants to have an honest conversation about it with close friends. One friend might accurately infer that the narrator's intention is to confess to a morally wrong action from his past, and respond in good faith. However, another friend might mistakenly infer that the narrator is trying to brag about his wife's accomplishments and to show what a good husband he is. The friend who misjudged the narrator's intention is then unlikely to arrive at the evaluation that the narrator desired (e.g., that he is blameworthy). In other words, audiences' accurate recognition of a narrator's intended meaning is important for avoiding accidental miscommunication. Further, to ensure transmission of their intended meanings (e.g., to meet their reputational goal of appearing morally good or bad), narrators should want to be clear and informative (i.e., have strong *informational* goals).

Note that according to this account, audiences do not need to accept the narrator's intended meaning upon recognizing it: they can disagree with the implied evaluations or reject the narrative as untrustworthy (Sperber et al., 2010; Mercier & Sperber, 2017). The possibility of being accidentally or intentionally misinformed is inherent to testimonial communication, leading audiences to guard against it with epistemic vigilance (Sperber et al., 2010). Audiences exercise epistemic vigilance when they do not blindly accept what they are told, but cautiously evaluate its trustworthiness and reliability, looking out for possible misinformation. Vigilance can lead to rejection when audiences become sufficiently suspicious about the testimony's veracity. Studies show that we exercise epistemic vigilance from a young age (Koenig, Clément, & Harris, 2004; Mazzarella & Pouscoulous, 2021; Mercier, 2020). For example, children as young as 5 years old use communicators' honest or deceptive intentions to believe or discount their testimony (Mascaro & Sperber, 2009; Shafto, Eaves, Navarro, & Perfors, 2012; Vanderbilt, Liu, & Heyman, 2011). Even younger children track and rely on communicators who have been informative in the past, as incompetent ones might unintentionally provide misinformation (e.g., Koenig et al., 2004; Jaswal & Neely, 2006; but see Jaswal 2010 and Vanderbilt et al., 2011).

Often, however, audiences are unable to verify a narrator's honest intentions (past or present) and must make informed guesses based on the testimony itself. Indeed, a crucial component of epistemic vigilance involves mind-reading the narrator to infer whether their intention is to be informative, helpful, and truthful (Harris et al., 2018; Mascaro & Sperber, 2009; Mazzarella & Vaccargiu, 2024; Sperber et al., 2010). How do audiences decide whether a testimony is likely to contain misinformation? Most past work addressing this question has focused on audiences' use of heuristics and cues that might signal misinformation. Studies on lie detection have examined the use of "tells" such as effort, body language, facial expressions, and speech disfluency to infer deceptive intent (e.g., "um…") (e.g., Ekman, Friesen, & O'sullivan, 1988; DePaulo et al., 2003; Vrij, 2008; Loy, Rohde, & Corley, 2018; Arciuli, Mallard, & Villar, 2010).

In this paper, we examine how audiences calibrate their trust based on goals inferred from the content of moral narratives. In particular, we test whether inferred reputational goals might increase distrust and discounting. We hypothesize that the extent to which audiences calibrate their trust depends on expectations about when narrators are likely to communicate misinformation. In moral narratives, narrators can transmit misleading information about their moral characters by self-enhancing or self-deprecating. Narrators might intentionally self-enhance, deceiving audiences to make themselves look better (e.g., by outright lying about a transgression or not mentioning it at all). Many studies suggest that enhancement is frequent in self-presentation (e.g., Alicke & Sedikides, 2009; Bolino, Kacmar, Turnley, & Gilstrap, 2008; Bourdage, Roulin, & Tarraf, 2018; Marsh & Tversky, 2004; Schlenker, 2003; Steinmetz, Sezer, & Sedikides, 2017). Self-enhancement can also be unintentional: a narrator could deceive themselves into thinking that they acted better than they actually did (e.g., Carlson, Maréchal, Oud, Fehr, & Crockett, 2020; Paulhus, 1991). Likewise, self-deprecation can occur both intentionally and unintentionally. People sometimes play down their accomplishments or status to avoid relational consequences, such as eliciting envy or appearing intimidating and arrogant (Arnett & Sidanius, 2018; MacGregor & Holmes, 2011; Roberts, Levine, & Sezer, 2021; Speer, 2019; Tice, Butler, Muraven, & Stillwell, 1995; Watling & Banerjee, 2007; Zell & Exline, 2010), or to appear more modest and trustworthy (Steinmetz et al., 2017). Others may hold overly harsh views of their own actions out of low self-esteem, guilt, and neuroticism (Owens, 1993; Owens, 1993; Luyten et al., 2007; Wolf et al., 2010). For

example, someone who doubts their positive deeds despite assurances from others, or someone who becomes very anxious about a minor transgression, might present an unduly worse image of their moral character.

As the most basic goal of the audience is to gain accurate information, they should be on the lookout for all possible causes of misinformation. Some reputational goals, however, are bound to be more likely than others. A large body of work shows that people want to appear morally good more often than they want to appear morally bad, although people do both (e.g., Alicke, 1985; Allison, Messick, & Goethals, 1989; Batson et al., 1999; Aquino & Reed II, 2002; Epley & Dunning, 2000; Shaw et al., 2014). This predicts that audiences may be more likely to expect positive than negative reputational goals in narrators. In addition, audiences may be more likely to expect narrators to self-enhance (look morally *better*) rather than self-deprecate (look morally *worse*). These expectations could interact as well, such that enhancement or deprecation could be suspected more when positive or negative reputational goals are inferred. In theory, even a narrator who is presenting themselves in a negative light might be self-enhancing (e.g., admitting to a transgression but minimizing how bad it was), and vice versa. However, since negative reputational goals are infrequent and costly, a narrator who is already presenting themselves as morally bad might be assumed to be more honest (or even self-deprecating) than a narrator presenting themselves as morally good. Alternatively, audiences might apply the same level of suspicion toward all moral narratives, regardless of the particular reputational goals they infer in narrators.

In the current study, we examine how audiences infer reputational and informational goals in moral narratives and how such inferences influence their comprehension and acceptance of narrators' intended meaning. We probed audience goal inferences in narratives from an experiment where narrators were incentivized to appear like a morally good or bad person while writing about a moral event from their own lives (Colombatto et al., 2024). We presented these moral narratives to two groups of audiences (Experiments 1 and 2) who were asked to report their inferences about narrators' goals (reputational and informational goal inference) as well as their actual evaluations of the narrators based on the contents of their narratives (moral character judgment). We first examined how accurately audiences detect narrators' actual reputational goals. Given the meaning-disambiguating function of goal recognition, narrators' reputational goals should be readily recognizable by audiences, and audiences' moral character judgments should align with what they inferred to be narrators' reputational goals. For example, an audience who thinks a narrator is trying to appear morally good should be more likely to view them as morally good. However, if audiences are suspicious toward particular reputational goals, their goal detection may cause judgments about moral character to systematically deviate from inferred reputational goals. For example, if audiences generally assume that narrators will self-enhance, they may always judge narrators to be worse than they are trying to appear. Importantly, if differences between inferred reputational goals and moral character judgments are driven by vigilance toward the narrator's trustworthiness, they should be accompanied by inferences of low informational goals.

2. Methods and materials

2.1. Overview of experiments

To explore audience inferences about narrators' goals, we first obtained narratives varying in reputational goals from a previous study of moral narratives (Colombatto et al., 2024) (Fig. 1, left). These narrators' ($n = 164$) were first instructed to write about a time in their lives when they did something morally questionable ('Original' narratives). This prompt was chosen with the aim of eliciting somewhat morally ambiguous narratives, so that the same actions could later be portrayed as morally worse or better. Further, we did not give any particular instruction with respect to reputational goals and expected narrators to represent their actions as neither clearly bad nor good. Three days later, narrators were invited to take part in a follow-up survey where they were incentivized to write about the same event again, two times – once trying to appear like a morally bad person ('Worst Person' narratives), and once trying to appear like a morally good person ('Best Person' narratives). This design resulted in sets of narratives describing the same events but with different reputational goals, allowing us to test a direct link between narrators' reputational goals and audience inferences of those goals.

Next, we ran two experiments testing perceptions of these narratives in naive audiences (Experiment 1: $n = 296$ and Experiment 2: n = 296). In Experiment 1, we conducted an initial test of goal detection in moral narratives: after reading about the narrative elicitation procedure, audience participants were shown all three narratives from each of several narrators and were asked to match each of the three narratives with its reputational goal ('Original', 'Best', 'Worst') (Fig. 1, middle).

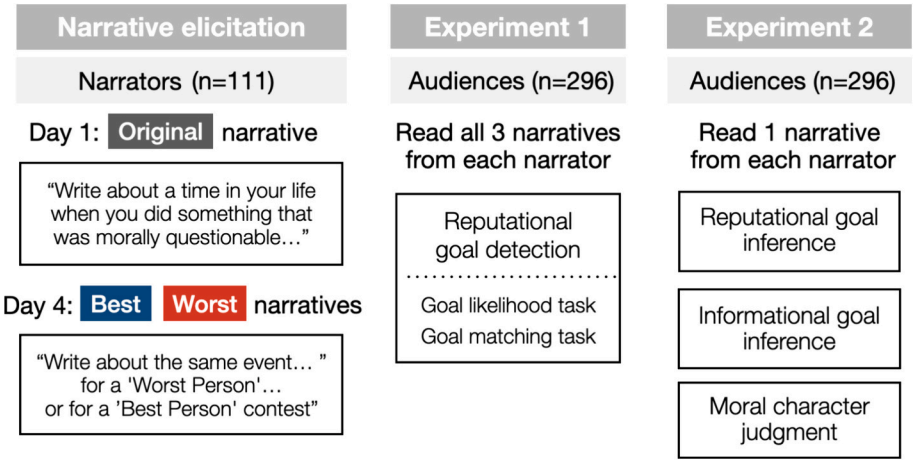

Fig. 1. Overview of experimental procedures. Left: Narrative elicitation in narrators from a previous study ($n = 111$). Narrators were asked to write about the same event three times, first with no particular instruction with respect to reputational goals ('Original'), and then with positive and negative reputational goals ('Best' and 'Worst' person). Middle: In Experiment 1, audiences (n = 296) read all three narratives from a given narrator and matched each with a reputational goal. Right: In Experiment 2, audiences (n = 296) read one narrative from each given narrator and made inferences about the narrators' reputational goals, informational goals, and moral character.

Since audiences in Experiment 1 were able to read all three narratives from each narrator and knew that narrators were given reputational goals, goal detection amounted to a three-alternative forced choice. In Experiment 2, we made the task more challenging by not informing audiences that narrators were prompted to write with positive or negative reputational goals. Further, audiences read only one narrative from each narrator. After reading a narrative, audience participants provided ratings for the narrator's reputational goals, informational goals, and moral character (Fig. 1, right). This allowed us to test how audiences detect goals in moral narratives, as well as how these impressions impact more global judgments of the narrators' moral characters.

2.2. Narrative elicitation

To investigate audience inferences about moral narratives written with different reputational goals, we used narratives collected as part of a previous study (Colombatto et al., 2024). These narratives were collected in a two-part study conducted over two days, four days apart. For this and all subsequent studies, all procedures were approved by the Yale University Institutional Review Board (#2000022385) and all participants provided informed consent and were compensated for their time. Note that all examples have been paraphrased to protect participant anonymity (Fig. 2).

On Day 1, 340 participants ('Narrators'; 153 female, 182 male; four nonbinary; one unspecified; mean age = 31.77) were recruited via Prolific (www.prolific.com). Narrators were instructed to *"write about a time in [their] life when [they] did something that was morally questionable: some people might think what [they] did was morally wrong, but others might think it wasn't so bad."* They were specifically asked to write about *"something that truly happened"*, avoiding illegal activities and omitting identifiable information. Next, narrators were asked to provide a short title for their narrative and were asked for permission for us to share their de-identified stories with other participants. Of the 340 participants, 53 were excluded for failure to follow the prompt, and the remaining 287 were invited to participate in the second part. Full instructions are available at https://osf.io/x3t5e.

On Day 4, narrators were invited to take part in a follow-up experiment; out of 287 participants who were invited, 225 (101 female, 120 male; three nonbinary; one unspecified; mean age = 32.05) completed this second survey. They were asked to write about the same event again and were told that these new stories would be entered into two contests – a Worst Person and a Best Person contest. Each participant wrote entries for both contests, one at a time, in a randomized order. Participants were told that judges would then later read their entries along with stories written by other participants and select the story that made "the author seem like the [worst/best] person". The winner of each contest would be awarded a $50 bonus. Narrators were thus asked to write the story from the beginning (as judges would not have access to their original stories), but now "trying to convince the judges that [they] were a very [immoral/moral] person." To ensure that narrators wrote about the same event across the three versions, they were further given these instructions: "You should feel free to include as few or as many details as you wish, but please remember it should be about the same story you wrote a few days ago, and it should still be based on events that truly happened." No other instructions were given about how narrators could try to appear like a Best or Worst Person. At the end of this two-part experiment, we excluded participants who reported not believing in the experimental manipulation ($N = 18$), those who indicated English was not their first language ($N = 3$), those who reported having encountered problems during the task ($N = 2$; e.g., "I misread the label and may have answered the opposite"), did not grant us permission to share their three stories ($N = 32$), or did not follow the prompt ($N = 38$; e.g., recounting different violations in each story, such as getting into a fight on the Original story, and then stealing in the Worst story; or not telling the story from the beginning, e.g. describing cheating on a test in the Original story, then just providing justifications in the following stories). This resulted in sets of three valid narratives from 132 narrators.

Next, we prepared these narratives to be shown to audiences. From the remaining 132 narratives, we removed those that were unfit to present to audiences due to excessive length, grammatical errors, de-anonymized materials, or leading statements (e.g., "This makes me a terrible person") – leaving a total of $N = 76$ narratives for Experiment 1. Because in Experiment 2 audiences read only one narrative from each narrator rather than all three, we had more flexibility in narrative selection, and we only excluded those that explicitly mentioned the prompt (e.g., "Something I did that was morally questionable is…") – leaving a total of $N = 111$ narratives for Experiment 2. Some example narratives can be found in Fig. 2.

2.3. Experiment 1: audience detection of reputational goals

2.3.1. Participants

A nationally representative sample of US participants was recruited from Prolific. We recruited 300 participants, with this pre-registered

Original	When I was in college, I cheated on a physics final. I had missed the final due to a medical emergency and had to take a make-up test at a later point. I had been doing ok in the class, but I always had difficulty in memorizing the equations. Because I was in a room by myself when I took the make-up test, I was able to look at a study sheet I made before, with all the equations written on it. I ended up doing much better on the test than I would have otherwise.
Best	In college, I had a medical emergency at end of one of my last semesters and missed several final exams. One of those exams was in physics - which I had put off taking until late in my degree because I have difficulty in remembering equations. Since then, I have been diagnosed with dyscalculia, but I didn't know that at the time. Anyway, I had been doing pretty well in the course, but I didn't have all the equations memorized. And I did make a study sheet with all the equations on it to prepare for the test, but I was so swamped with catching up that I couldn't study as much as I usually would. Fast forward to the make-up test: the professor left me alone in a room to take it. Even though I felt extremely guilty about it, I used the study sheet I made. I had never cheated on a test or an assignment prior to that and the fact that I got away with it only made my guilt worse.
Worst	I missed a final exam in physics when I was at university. The professor let me take a make-up test and gave me the test to take in a room by myself. I didn't like physics very much - it wasn't my major and I was taking it because it was a required "general education" requirement. I hadn't spent much time trying to memorize the equations before the test. Once I was left alone, I used a study sheet of equations that I had made before, even though I wasn't supposed to. I didn't get caught and I did better than I would have done otherwise by cheating.

Fig. 2. Representative narratives. Original, Best, and Worst narratives from one narrator (paraphrased to preserve anonymity). More examples are included in Supplemental Materials.

sample size chosen as the minimum to achieve a representative sample on Prolific. Of 325 participants who completed the study, we excluded according to our pre-registered criteria those who reported having encountered problems ($N = 5$), failed to answer our debriefing questions sensibly ($N = 6$), or selected the wrong option in a comprehension check (N = 18; see Experimental procedures for full text). This left a valid sample of $N = 296$ participants (138 female, 140 male; two nonbinary; one gender neutral; 15 unspecified; mean age = 44.21).

2.3.2. Experimental procedures

Audiences were first given all details about how the narrators were prompted. They were informed that narrators were other participants on Prolific who were asked to write about a true event from their own lives where they did something morally questionable, and that they wrote the same story three times (first, as they remembered the events; second, while trying to appear like a very morally good person to win a 'Best Person' contest for a bonus, and third, while trying to appear like a very morally bad person to win a 'Worst Person' contest) (Fig. 3, left).

To ensure that participants understood the structure of narrators' task, they completed a comprehension check where they had to select what the narrators were asked to do ("Write about a time in their lives where they did something nice for others."; "Write about the same story three times: one where they appear to be good, one where they appear to be bad, one where they just tell us what happened."; or "Write about three different stories: one where they did something good, one where they did something bad, and one where they did something neither good nor bad."). Audiences were then shown the three narratives from the same narrator (labeled "Story 1", "Story 2", "Story 3") in randomized order and on separate screens (Fig. 3, left). Participants were allowed to move on to the next narrative only after 10 s to ensure they fully read each story.

After reading all three narratives, participants were asked to indicate how likely that story was to be the 'Original', 'Best Person', and 'Worst Person' story, on a scale of 'Not at all likely' to 'Very likely' (likelihood ratings) (Fig. 3, right). Audience participants made these ratings for each narrative on separate screens, with all three versions shown at the top of the page as a reminder. Audiences were also asked to make a forced choice by dragging the story labels into 'Original', 'Best Person', and 'Worst Person' boxes (matching task). Again, all three stories were displayed on the page, allowing participants to read them again for comparison. Finally, audiences made judgments about the narrators' likability, trustworthiness, and moral character (not analyzed in this paper; see Colombatto et al., 2024). Each audience participant completed this task for sets of narratives from 5 randomly selected narrators, followed by some demographic questions.

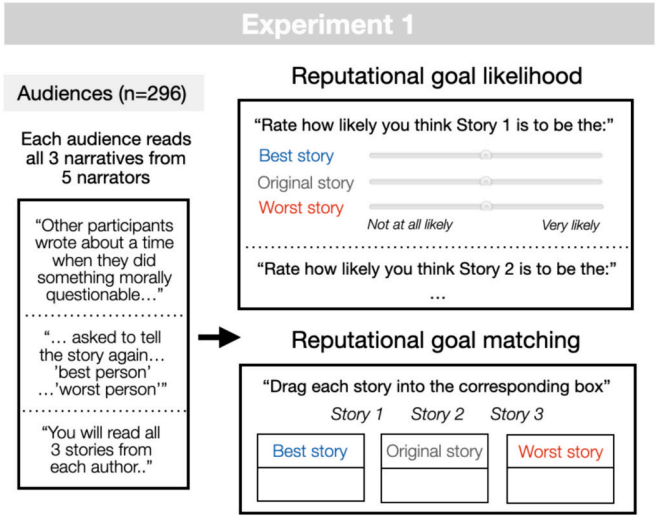

Fig. 3. Procedures for Experiment 1. Left: Audiences were informed about the instructions narrators received. Right: Likelihood ratings and goal matching tasks used to probe reputational goal detection.

2.4. Experiment 2: audience detection of goals and moral character judgments

2.4.1. Participants

A nationally representative sample of US participants was recruited from Prolific. Of 300 participants who completed the study, we excluded according to our pre-registered criteria those who failed to answer our debriefing questions sensibly ($N = 4$), and no participants reported having encountered problems. This left a valid sample of $N = 296$ participants (136 female, 139 male; three nonbinary; 18 unspecified; mean age = 45.20).

2.4.2. Experimental procedures

In contrast to Experiment 1, audience members in Experiment 2 read only one narrative from each narrator (Fig. 4). Each audience member read 12 narratives (4 'Original', 4 'Best', 4 'Worst', each written by a different narrator and shown in a randomized order). Further, we no longer informed audience participants about the reputational goal manipulation, but rather simply informed them that they would read narratives from other participants: *"In this study, you are going to read some stories about moral situations written by other participants on Prolific. The participants […] could write about any event they wanted to as long as it was something that truly happened to them."*

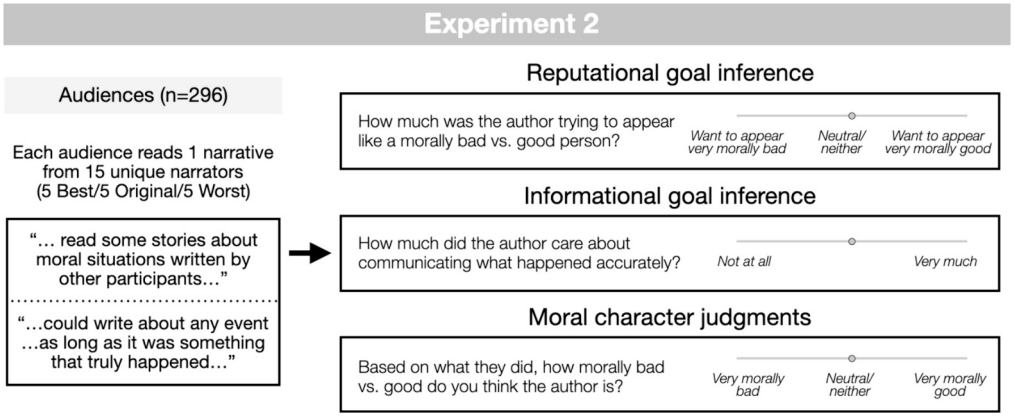

Fig. 4. Procedures for Experiment 2. Audiences only read one narrative from each narrator ($N = 111$). Left: Audiences were only told that narrators wrote about a true event involving some moral situation. Right: Questions probing audience inferences about the narrator's reputational goal, informational goal, and moral character.

After reading each story for at least 10 s, audience participants answered three questions probing reputational goal inference ("How much was the author trying to appear like a morally bad vs. good person?", with slider labels "Want to appear very morally bad", "Neutral/neither", and "Want to appear very morally good"); informational goal inference ("How much did the author care about communicating what happened accurately?", with slider labels "Not at all" and "Very much"), and moral character judgment ("Based on what they did, how morally bad vs. good do you think the author is?", with slider labels "Very morally bad", "Neutral/neither", and "Very morally good") (Fig. 4, right).

3. Results

3.1. Audiences accurately detect narrators' reputational goals

Using three tasks (likelihood, matching, inference) across two experiments, we examined how accurately audiences detect narrators' reputational goals. In Experiment 1, audiences were told that each narrator wrote three narratives with different reputational goals. After reading all three narratives from the same narrator, audiences rated the likelihood of each narrative corresponding to the three reputational goals (Best/Original/Worst) (Fig. 5, left). They additionally matched each of the narratives to one of the three goals (Fig. 5, middle). In Experiment 2, audiences were not told about the reputational goal manipulation and only saw one narrative from each narrator. Audiences then indicated their inferences about whether the narrator wanted to appear very morally bad or good (Fig. 5, right). Goal detection accuracy was high across all tasks.

Following our pre-registered plan, likelihood ratings (Fig. 5, left) were analyzed using a linear mixed effect model with the narrator's incentivized reputational goal ('Original', 'Best', 'Worst') and audience rating prompts ('Original', 'Best', 'Worst') as interacting fixed effects (dummy coded with Original as reference), as well as audience participants as random effects. (Our pre-registered analysis included narrative nested within narrators as random effects as well, but this resulted in a singular fit, and so we report a model with a simplified random effects structure.) As would be expected if audiences consistently discriminated between narratives written with different goals, there was an interaction between audience rating prompts and narrator's incentivized goals ($F(4, 13{,}004) = 1076.49$, $p < .001$). There was a significant main effect of audience prompt ($F(2, 13{,}004) = 60.03$, $p < .001$), but not of narrators' goals ($F(2,13{,}004) = 2.18$, $p = .113$).

Post-hoc tests (Bonferroni corrected) confirmed that audiences were highly accurate at reputational goal detection: for each rating prompt, likelihood ratings were highest for the prompts that matched the narrator's incentivized goal. Best Person narratives received higher Best likelihood ratings compared to Original ($B = 34.95$, $SE = 1.22$, $t(13004) = 28.63$, $p < .001$) and Worst ratings ($B = 52.21$, $SE = 1.22$, $t(13004) = 42.76$, $p < .001$). Similarly, Original narratives received higher Original likelihood ratings compared to Best ($B = 35.75$, $SE = 1.22$, $t(13004) = 29.28$, $p < .001$) and Worst ratings ($B = 37.28$, $SE = 1.22$, $t(13004) = 30.54$, $p < .001$). Likewise, Worst Person narratives received higher Worst likelihood ratings compared to Original ($B = 31.40$, $SE = 1.22$, $t(13004) = 25.72$, $p < .001$) and Best ratings ($B = 43.11$, $SE = 1.22$, $t(13004) = 35.30$, $p < .001$).

Best and Worst Person narratives were unlikely to be confused as the other – Best narratives received higher Original compared to Worst likelihood ratings ($B = 17.26$, $SE = 1.22$, $t(13004) = 14.14$, $p < .001$), and Worst narratives received higher Original compared to Best likelihood ratings ($B = 11.71$ $SE = 1.22$, $t(13004) = 9.59$, $p < .001$). Further, Original narratives received similar Best and Worst likelihood ratings ($B = 1.53$, $SE = 1.22$, $t(13004) = 1.25$, $p = .211$).

Matched goals (Experiment 1) were analyzed using a multinomial logistic regression, with the narrator's incentivized reputational goal as a fixed effect, and narrators and audiences as random effects (Fig. 5, middle). As with the likelihood ratings, audiences were very accurate at detecting narrators' reputational goals. There was a significant effect of narrators' goals on the audience's matching choices ($\chi 2(4) = 2185.2$, $p < .001$). Narratives written with Best Person goals were much more likely to be matched by audiences as Best than as Original (log odds, $B = 2.40$, $SE = 0.12$, $t(6) = 20.66$, $p < .001$) or Worst ($B = 2.89$, $SE = 0.12$, $t(6) = 23.58$, $p < .001$), Original narratives were more likely to be matched as Original than Best ($B = 2.28$, $SE = 0.11$, $t(6) = 20.27$, $p < .001$) or Worst ($B = 1.93$, $SE = 0.11$, $t(6) = 17.68$, $p < .001$), and Worst Person narratives were more likely to be matched as Worst than Best ($B = 2.55$, $SE = 0.12$, $t(6) = 21.86$, $p < .001$) or Original ($B = 2.24$, $SE = 0.11$, $t(6) = 19.82$, $p < .001$).

Similar to the likelihood ratings, Best Person narratives were more likely to be matched as Original than Worst (log odds, $B = 0.49$, $SE = 0.11$, $t(6) = 4.30$, $p = .005$), and Worst Person narratives were more likely to be matched as Original than Best ($B = 0.31$, $SE = 0.11$, $t(6) =$

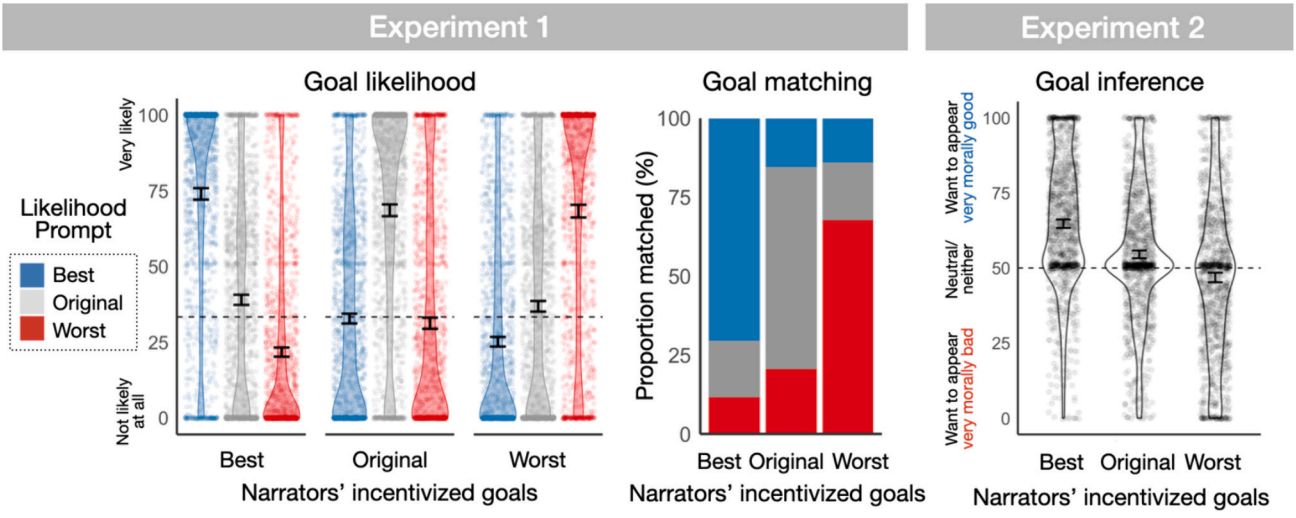

Fig. 5. Audience inferences about narrators' reputational goals across Experiments 1 and 2. Results are presented according to narrators' actual goals on the x-axis (i.e., whether their narrative was generated in response to Best/Worst Person contest, or was the Original narrative). Left: likelihood ratings ("Rate how likely you think the Story is the [Best/Original/Worst]"). Middle: proportion of each narrative type matched as Best/Original/Worst. Right: goal inference ratings from Experiment 2. Error bars are mean +/− 95% confidence intervals subtracting out the shared variance.

2.80, $p = .031$). However, Original narratives were more likely to be Worst than Best (B = 0.35, SE = 0.11, t(6) = 3.30, $p = .016$).

Audience inferences about reputational goals in Experiment 2 were accurate as well, even though by design audiences could not compare the three narratives from each narrator (Fig. 5, right). Goal inference ratings were analyzed using a linear mixed effects model with narrative type as a fixed effect and audiences and narratives nested within narrators as random effects. There was a main effect of narrator goal (F(2, 206) = 81.73, $p < .001$) where narrators of Best Person narratives were inferred as wanting to appear morally good more than narrators of Original (B = 10.08, SE = 1.39, t(210) = 7.26, p < .001) and Worst Person narratives (B = 17.73, SE = 1.39, t(210) = 12.74, p < .001), and narrators of Original Person narratives were inferred as wanting to appear morally good more than narrators of Worst Person narratives (B = 7.65, SE = 1.39, t(210) = 5.51, p < .001).

Note that unlike in Experiment 1, audiences in Experiment 2 did not know that narrators were given any goals at all. This allowed us to explore what kinds of reputational goals audiences expect narrators to have. Goal inference ratings showed that audiences were more likely to infer positive reputational goals (all reputational goal ratings >50: t(3549) = 13.44, p < .001; M = 55.37, CI[54.59 56.16]) (Fig. 5, right). Reputational goal ratings for both Best and Original narratives tended toward "want to appear morally good" (for Best >50: t(1183) = 23.30, p < .001; M = 64.73, CI[63.49 65.97]; for Original >50: t(1182) = 7.51, p < .001; M = 54.53, CI[53.35 55.71]). Since audiences in Experiment 2 did not know that narrators had been assigned specific reputational goals, this suggests that audiences may assume narrators are more likely to be driven by positive reputational goals.

3.2. Audiences discount more when positive reputational goals are inferred

To summarize so far, we found across two experiments that audiences accurately detect narrators' reputational goals, and audiences expect narrators to have positive reputational goals. In the following analyses, we investigate each of these findings further, asking whether the effects reflect strategies taken by narrators to gain audience trust, and whether such strategies are successful.

In Experiment 2, audiences were asked to judge the narrator's moral character. Comparing audiences' inferred reputational goals and actual character judgments allows us to examine whether audiences' impressions of narrators matched the impressions audiences thought narrators desired. Indeed, inferred reputational goal ratings and character judgments were highly correlated (Fig. 6, left, $r = 0.51$, t(3548) = 35.61, $p < .001$). These results are consistent with the explanation that audiences' judgments are likely to align with a narrator's reputational goal when they accurately infer it.

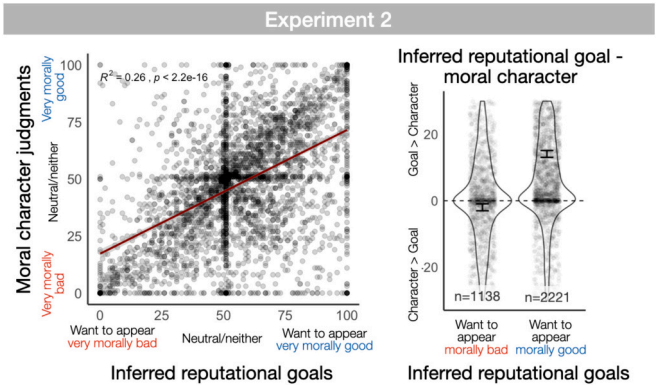

Fig. 6. Comparison between audience inferences of narrators' reputational goals vs. their judgments of narrators' character. Left: correlation between inferred reputational goal and character judgments. Right: the amount of adjustment (deviation of character judgments from inferred goals) split by whether the inferred goal was positive (want to appear morally good) or negative (want to appear morally bad). Error bars are mean +/− 95% confidence intervals subtracting out the shared variance.

Character judgments, however, were not perfectly aligned with inferred reputational goals. Given our previous result that audiences expect narrators' reputational goals to be more positive than negative, we tested whether they also expect narrators to always self-*enhance* (i.e., try to appear better than they actually are). Note that inferred positive and negative reputational goals refer to inferences about whether narrators are trying to appear morally good or bad, not whether they are trying to appear morally *better* or *worse*. Audiences could in principle infer either a positive or negative reputational goal and infer that the narrator is truthfully representing their own moral character, trying to appear better, or trying to appear worse. If audiences generally expect self-enhancement, they should downwardly adjust their character judgments (i.e., judge the narrator to be morally worse than they are trying to appear) regardless of the inferred reputational goal (i.e., positive or negative).

We examined the difference between audience's inferences of narrators' reputational goals and their character judgments, split by the direction of the inferred goal (positive: ratings >50, corresponding to responses toward "want to appear very morally good" or negative: ratings <50, corresponding to responses toward "want to appear very morally bad") (Fig. 6, right). For positive inferred goals, audiences judged narrators' actual moral characters to be worse than they are trying to present ((goal - actual) for positive: M = 14.1, SD = 0.52, CI [13.07 15.13], t(2220) = 26.88, $p < .001$). In contrast, when the inferred goal was negative, audiences judged the narrator to be slightly *better* than they are trying to appear ((goal – actual) for negative: M = -3.69, SD = 0.53, CI[−3.02–0.92], t(1328) = −3.69, p < .001). This suggests that audiences expect narrators' goals to be *exaggerations* in either positive or negative directions (i.e., trying to appear better than they are if they have a positive goal and trying to appear worse than they are if they have a negative goal).

Alternatively, this result could reflect a floor and ceiling effect or different uses of scales for the inferred reputational goal and moral character judgment questions. However, the discrepancy between inferred reputational goals and character judgments was also much larger when positive goals were inferred (t(2740) = 21.2, p < .001). This suggests that audiences disproportionately discount inferred positive reputational goals as self-enhancements. This result is consistent with audiences being more likely to expect positive, self-*enhancing* goals than negative, self-*deprecating* goals. Even when negative goals are detected, audiences think most narrators are unlikely to want to appear much worse than they actually are.

3.3. Achieving positive reputational goals requires audience inference of high informational goals

We next tested the possibility that audiences' inferences of narrators' informational goals impacts their trust in narrators. In our main preregistered analysis, we looked at the interactive effects of inferred informational and reputational goals on actual character judgments using another linear mixed effects model (Character ~ Reputational goal * Informational goal + same random effects as before) (Fig. 7). While inferred informational goals alone did not have an effect on character judgments (B = 0.03, SE = 0.03, CI[−0.01 0.09], t(3388) = 1.18, $p = .2$), there was a significant interactive effect of informational and reputational goals (B = 0.003, SE = 0.0049, CI[0.00 0.00], t(3359) = 6.48, p < .001) and a main effect of reputational goal (B = 0.16, SE = 0.04, CI [0.09 0.24], t(3432) = 4.29, p < .001). Comparing a model without the interaction also confirmed that the full model is better ($\chi^2(1) = 41.75$, p < .001).

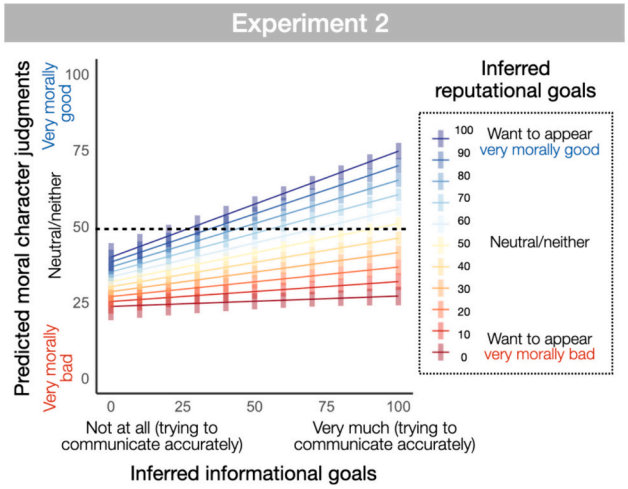

Fig. 7. Interactive effect of inferred informational and reputational goals on moral impressions. Model outputs for the interactive effects of inferred informational and reputational goals on moral character judgments. Y-axis shows predicted actual reputation ratings, x-axis shows inferred informational goal ratings, and colors indicate levels of inferred reputational goal ratings.

This interaction effect shows that if a narrator is inferred to have a negative reputational goal, whether they are inferred as trying to be informative does not matter much for character judgment – audiences will judge them as similarly morally bad. In contrast, if a narrator is inferred to have a positive reputational goal, then appearing informative is crucial: the presence of both goals is necessary for narrators to be judged as morally good. An informative narrator who is perceived as trying to appear good is judged to be (almost) as good as their inferred goal reputation, but an uninformative narrator with an inferred positive goal is judged to be much morally worse than their inferred goal. Importantly, the interaction between informational and reputational goals also suggests that the deviation of audiences' moral character judgments from inferred reputational goals is driven by suspicion that narrators are not being truthful.

3.4. Narrators try harder to appear informative when reputational goals are positive

To summarize the results so far: first, narrators' reputational goals (and positive ones in particular) are easily detectable. One benefit of making these goals detectable for audiences is that their beliefs about narrators' actual reputations align closely with what they think narrators want them to believe. However, we also find evidence that audiences do not completely accept narrators' reputational goals: they expect narrators to generally want to appear better than they are. Importantly, inferred reputational goals interact with inferred informational goals. Being seen as wanting to be informative is particularly important for trust when the audience infers a positive reputational goal.

A possible implication of this last result is that audiences expect narrators with positive reputational goals to be less truthful. This predicts that inferred positive goals might lead to perceptions of low informational goals. In other words, while our previous analysis assumed informational and reputational goals to be independent, there might be a relationship between them. To test this hypothesis, we returned to analyzing inferred informational goals as a function of what reputational goals narrators were assigned to have (Best/Worst/Original). Surprisingly, when we look at inferences about informational goals ("did the narrator care about communicating what happened accurately") by narrative type, audiences rated Best narratives as having the highest informational goals (LMEM predicting goal ratings, effect of narrative type: $F(2,194.12) = 11.09$, $p < .001$; contrasts of Best vs. Original: $B = 4.26$, $SE = 1.01$, $t(194.38) = 3.95$, $p < .001$, Best vs. Worst: $B = 4.52$, $SE = 1.08$, $t(193.19) = 3.49$, $p < .001$; Fig. 8, left).

As this was not predicted, we conducted further exploratory analyses to understand why Best narratives were perceived as most informative. One difference we observed across narratives is that narrators tended to write more for the Best narratives, followed by Worst, then Original (word count for Best: $M = 164.43$ words, $SD = 88.84$; Original: $M = 96.6$, $SD = 43.1$; Worst: $M = 135.6$, $SD = 81.68$; Wilcoxon signed-rank tests, two-tailed for Best vs. Worst: $z = 4.41$, $p < .001$; Best vs. Original: $z = 7.72$, $p < .001$; Original vs. Worst: $z = 5.50$, $p < .001$) (Fig. 8, middle). Word length was in turn positively correlated with inferred informational goals ($r = 0.37$, $p < .001$) (Fig. 8, right). The relationship between informational and reputational goals was positive even when controlling for word count and including Best narratives only (LMEM predicting informational goal ratings, effect of reputational goal: $B = 0.15$, $CI[0.09, 0.21]$, $F(1, 1004) = 23.21$, $p < .001$; effect of word length: $B = 0.19$, $CI[0.14, 0.25]$, $F(1,139) = 46.85$, $p < .001$; interaction between word length and reputational goal: $B = -0.04$, $CI[-0.09, 0.00]$, $F(1,1034) = 3.2$, $p = .07$). These results suggest that narrators might have very accurate understandings of audiences' likely inferences and put more effort into writing the Best narratives so that they will be viewed as having high informational goals and be accepted.

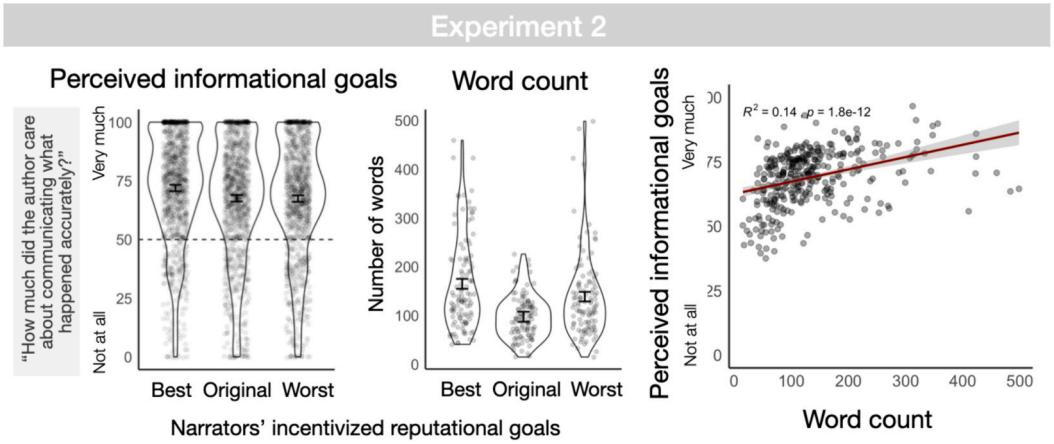

Fig. 8. Inferences about narrators' informational goals. Left: audience's inferences of narrators' informational goals, arranged by the narrators' instructed reputational goals. Middle: word length for each type of instructed reputational goal. Right: correlation between word length and inferred informational goals. Error bars are mean +/− 95% confidence intervals subtracting out the shared variance.

4. Discussion

People frequently give testimony about their own moral character by telling moral narratives. Because narrators often want to appear morally better than they are, audiences need to exercise epistemic vigilance (Sperber et al., 2010). Often, as audiences, we have to make inferences about narrators' goals (e.g., informational and reputational goals) based on the content of their testimony alone. How good are audiences at detecting narrators' positive and negative reputational goals? What goals do audiences expect in narrators? How much do audiences accept or discount what they infer narrators as wanting to communicate? And how do inferred informational and reputational goals affect audiences' acceptance of moral narratives?

Across two experiments, we find that audiences detect narrators' incentivized reputational goals with high accuracy. When audiences did not know that the narrators were *given* particular reputational goals (Experiment 2), they tended to infer positive goals in narrators. While inferred reputational goals and moral character judgments in audiences were highly correlated, there were systematic deviations. If audiences inferred positive reputational goals, narrators' moral characters were judged to be *worse* than their inferred goal reputations. In contrast, when narrators' inferred reputational goals were negative, audiences judged them to be slightly better than their goal reputations. Further, there was an asymmetry such that this deviation between moral character judgments and inferred reputational goals was greater for inferred positive goals.

Importantly, we also found an interactive effect of inferred informational and reputational goals on moral character judgments: the deviation between inferred reputational goals and moral character was greatest when low informational goals were inferred, only for positive reputational goals. This result suggests that audiences adjusted from what they inferred narrators as wanting them to believe out of suspicion toward the narrator's trustworthiness. Finally, we found indirect evidence that narrators might have written their narratives with accurate predictions about audiences' likely inferences in mind: narrators wrote longer narratives when they had positive reputational goals, which resulted in audience inferences of higher informational goals. In other words, narrators may have tried harder to appear informative when their reputational goals were positive because they knew audiences would be more suspicious of positive reputational goals. Another possibility that we cannot rule out here is that narrators with positive reputational goals were successful at appearing morally good, and perceptions of goodness increased perceptions of communicative competence (e.g., Stellar & Willer, 2018).

The asymmetrical discounting between positive and negative reputational goals is consistent with the idea that audiences may expect that narrators are more likely to want to appear morally good than bad. A large literature suggests that people generally want to appear good to others, and to themselves (e.g., Steinmetz et al., 2017; Schlenker, 2003; Marsh & Tversky, 2004; Bourdage et al., 2018; Alicke, 1985; Allison et al., 1989; Batson et al., 1999; Aquino & Reed II, 2002; Epley & Dunning, 2000; Shaw et al., 2014; Paulhus, 1991; Sherman & Cohen, 2006; Carlson et al., 2020). From an audience perspective, a narrator who seems to have a positive reputational goal could be presenting themselves as morally good because they 1. genuinely and truthfully believe themselves to be good, 2. believe themselves to be bad but deceptively want to look better, or 3. falsely believe themselves to be better than they are (e.g., due to self-deception). Similarly, a narrator who appears to have a negative reputational goal could 1. genuinely and truthfully believe themselves to be bad, 2. believe themselves to be morally good but deceptively want to look worse, or 3. falsely believe themselves to be worse than they are. Given a prior that people tend to want to appear good, audiences might infer that option 2 and 3 are likely when the narrator's inferred reputational goal is positive but far less likely when the inferred reputational goal is negative. In other words, audiences are less suspicious toward negative reputational goals because they are less likely to lead to misinformation.

At the same time, audiences did not automatically doubt all positive reputational goals. The discounting we observed with positive reputational goals was small and hinged on additional inferences of lower informational goals. Discounting was minimal even though informational and reputational goal inference questions were always asked before moral character judgments, likely priming audiences to think about the goals. This may be explained by audiences taking a 'stance of trust', where they remain skeptical and wary without actively disbelieving a testimony (Levine, Mikhail, & Leslie, 2018; Mascaro, Morin, & Sperber, 2017; Sperber et al., 2010). Whether audiences are justified in maintaining this amount of skepticism should be tested by examining how much narrators deceive – for example, by looking more precisely at whether audience expectations of narrators' rates of self-enhancement match how much narrators spontaneously self-enhance. In a sense, audiences were "correct" in the current study to not suspect narrators at high rates – our narrators were instructed to only write about things that actually happened. The different narratives they produced did not contain lies, but rather, reflected changes in framing, where narrators changed which details were included, emphasized, and worded.

Another possible explanation for why we did not see bigger deviations between inferred reputational goals and moral character judgments is that even when audiences distrusted a narrator, they had trouble rejecting the narrative they had entertained. Such belief perseverance or failure to course correct in the face of known misinformation has been frequently documented in memory (e.g., Anderson, 1983; Green & Donahue, 2011; Marsh, Meade, & Roediger III, 2003). In Green and Donahue (2011), for example, audiences read a story and then were told the story was false. Upon finding out that the author of the story intended deception, audiences judged them to be immoral, and discounted some parts of the story, but continued to believe many of the facts revealed to be false. One explanation offered by Green & Donahue is that even when deception is revealed, it can be difficult to construct an alternative or know exactly how much to correct as well as knowing precisely which details to discount (e.g., if you realize someone is trying to appear like a morally good person, do you decide that they are now a "bad" person, or just a slightly less good person).

Intriguingly, there are proposals that the format of narratives may make discounting particularly difficult for audiences. Narratives are thought to "transport" and immerse readers into a narrative world, making them less aware of their real surroundings (Green & Brock, 2000). Fraser (2021) goes beyond transportation to argue that the structure of narratives creates a deeper dependence than other forms of communication, shaping what audiences infer and pay attention to. Such perspectival dependence could lead to difficulty by audiences to form a belief that is different from the one being presented to them, even if they are able to recognize the narrators' intentions and exercise vigilance toward them. While the narratives elicited in the current study may not be as immersive as fictional stories typically used in studies on transportation, they are similarly rich in detail, providing particular, personal, coherent, and concrete information, from a single perspective (Fig. 2, Table S1 in Supplemental Materials). A related possibility is that suspicion and discounting decreases with narrators' ability to mask their own goals (Bietti et al., 2019; Reboul, 2011, 2017; Vallauri et al., 2020). According to this account, the more audiences are made to feel like they arrived at a belief themselves (rather than through the narrators' machinations), the more audiences are likely to align to the narrators' intended belief. In this study, we found that audiences' beliefs about moral character aligned with narrators' inferred reputational goals in spite of accurate goal recognition. An open question is whether audiences discount even less when narrators' goals are harder to detect (e.g., when narratives are particularly engaging and effective at transporting the audience).

To the extent that there was any perception of low informational goals alongside negative reputational goals, the discounting went in the opposite direction: audiences perceived narrators' moral characters to

be slightly better than their inferred goal reputations. One possibility is that audiences discounted negative reputational goals because they found them implausible. However, a substantial proportion of audiences perceived negative reputational goals, even when they were not informed of the goal manipulation in narrators (Experiment 2; Fig. 5, right). While less studied, there are documented situations in which people earnestly present themselves in a negative light. For example, feelings of guilt, remorse, and shame might lead someone to apologize or confess to a wrongdoing, seeking to portray themselves negatively to signal that they understand the severity of their transgression (Exline, Deshea, & Holeman, 2007; Kelley & Waldron, 2005; Knight, 2018). Similarly, people might seek to portray their own transgressions in a negative light while talking to others in order to make sense of the severity of their actions (Rimé et al., 1998; Pasupathi, McLean, & Weeks, 2009; Mansfield, Pasupathi & McLean, 2015; Mansfield, McLean, & Lilgendahl, 2010). Further, since narrators in the current experiment disclosed minor, everyday wrongdoings (e.g., cheating, stealing, lying, as opposed to murder; Fig. S1), and many emphasized personal growth following their wrongdoings, we think it unlikely that audiences found negative reputational goals to be completely implausible in narrators.

Another open question is whether inferences about informational and reputational goals are dependent on each other in any way. We interpreted the finding that narrators with inferred positive reputational goals are seen as more informative as driven by the perception of narrators' increased effort in writing these narratives. This would be a rational strategy given that perceptions of self-enhancement lead to negative judgments (e.g., Schlenker & Leary, 1982; Van Damme, Hoorens, & Sedikides, 2016; Vonk, 1999). Further, studies on people's lay theories about altruistic acts suggest that narrators should expect audiences to be suspicious toward positive reputational goals. People often assume self-interested motives in prosocial acts, and inferred self-interested motives lead to harsher moral judgments (Carlson, Bigman, Gray, Ferguson, & Crockett, 2022; Carlson & Zaki, 2022; Newman & Cain, 2014; Berman & Silver, 2022; Lin-Healy & Small, 2013; but see also Zhang et al., 2021, Kawamura, Sasaki, & Kusumi, 2022 for cultural differences). On the other hand, the pronounced discounting for positive reputational goals suggests that the perception of positive goals should have led to perceptions of low informational goals, but this is not what we find. One possibility is that we may not have been able to detect this relationship because narrators' efforts to appear informative were overwhelming. Another possibility is that because our informational goal question probed perceptions of accuracy, not truth or trustworthiness per se, the resulting ratings did not capture audiences' suspicions. Future work should directly test whether perceptions of positive reputational goals also raise audience suspicion.

Together, these investigations of goal inference in moral narratives demonstrate how audiences' inferences about narrators' goals can both support and hinder the acceptance of the narrators' intended meanings. Accurate goal recognition increases the chance that audiences will make judgments intended by narrators, but it can also lead to vigilance about accuracy, depending on their expectations about particular goals. An important future direction is to compare the relationship between perceptions of informational and reputational goals and moral character judgment across cultures, as norms and expectations about honesty and self-presentation can vary greatly (e.g., Fu, Xu, Cameron, Heyman, & Lee, 2007; Griffin & Bender, 2019; Lee, Xu, Fu, Cameron, & Chen, 2001; Rui & Stefanone, 2013). Further, audiences' inferences and prior assumptions about narrators' informational and reputational goals may change if the narrator is providing a story about a third party, rather than themselves. Future work should also examine differences in individual narrators' ability to achieve their reputational goals, as well as how much individual audiences suspect and distrust narrators. More broadly, by combining insights from narrative theory, moral psychology, impression management, and linguistics, we show how comprehension and trust can be studied in complex, naturalistic forms of communication, such as narratives.

CRediT authorship contribution statement

J.K.: Conceptualization, Methodology, Formal analysis, Visualization, Writing - Original draft; C.C: Conceptualization, Methodology, Formal analysis, Visualization, Writing - Review & Editing; M.J.C: Conceptualization, Supervision, Writing - Review & Editing, Funding acquisition.

Declaration of competing interest

None.

Data availability

All the data and code are available on https://osf.io/x3t5e. Per our IRB, we cannot include the original narratives written by participants.

Acknowledgements

We thank Danny Rodriguez for his help collecting and cleaning the data, as well as members of the Crockett Lab for helpful discussions of this work. This work was supported by the John Templeton Foundation grant on How Stories Inspire Virtuous Behavior (#61495 & #62816) awarded to MJC.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105865.

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.

Alicke, M. D., & Sedikides, C. (2009). Self-enhancement and self-protection: What they are and what they do. *European Review of Social Psychology, 20*(1), 1–48.

Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On being better but not smarter than others: The Muhammad Ali effect. *Social Cognition, 7*(3), 275–295.

Anderson, C. A. (1983). Abstract and concrete data in the perseverance of social theories: When weak data lead to unshakeable beliefs. *Journal of Experimental Social Psychology, 19*(2), 93–108.

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

Arciuli, J., Mallard, D., & Villar, G. (2010). "Um, I can tell you're lying": Linguistic markers of deception versus truth-telling in speech. *Applied PsychoLinguistics, 31*(3), 397–411.

Arnett, R. D., & Sidanius, J. (2018). Sacrificing status for social harmony: Concealing relatively high status identities from one's peers. *Organizational Behavior and Human Decision Processes, 147*, 108–126.

Aronowitz, S., & Lombrozo, T. (2020). Experiential explanation. *Topics in Cognitive Science, 12*(4), 1321–1336.

Banerjee, R., Heyman, G. D., & Lee, K. (2020). The development of moral self-presentation. In *The Oxford handbook of moral development: An interdisciplinary perspective* (pp. 92–109).

Batson, C. D., Thompson, E. R., Seuferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525.

Baumeister, R. F., Zhang, L., & Vohs, K. D. (2004). Gossip as cultural learning. *Review of General Psychology, 8*(2), 111–121.

Berman, J. Z., & Silver, I. (2022). Prosocial behavior and reputation: When does doing good lead to looking good? *Current Opinion in Psychology, 43*, 102–107.

Bietti, L. M., Tilston, O., & Bangerter, A. (2019). Storytelling as adaptive collective sensemaking. *Topics in Cognitive Science, 11*(4), 710–732.

Bolino, M. C., Kacmar, K. M., Turnley, W. H., & Gilstrap, J. B. (2008). A multi-level review of impression management motives and behaviors. *Journal of Management, 34*(6), 1080–1109.

Bonalumi, F., Mahr, J. B., Marie, P., & Pouscoulous, N. (2023). Beyond the implicit/explicit dichotomy: The pragmatics of plausible deniability. *Review of Philosophy and Psychology*, 1–23.

Bourdage, J. S., Roulin, N., & Tarraf, R. (2018). "I (might be) just that good": Honest and deceptive impression management in employment interviews. *Personnel Psychology, 71*(4), 597–632.

Brown, P., & Levinson, S. C. (1978). Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction* (pp. 56–311). Cambridge University Press.

Carlson, R. W., Bigman, Y. E., Gray, K., Ferguson, M. J., & Crockett, M. J. (2022). How inferred motives shape moral judgements. *Nature Reviews Psychology, 1*(8), 468–478.

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications, 11*(1), 2100.

Carlson, R. W., & Zaki, J. (2022). Belief in altruistic motives predicts prosocial actions and inferences. *Psychological Reports, 125*(4), 2191–2212.

Colombatto, C., Conway, P., De Brigard, F., Hofmann, W., Sinnott-Armstrong, W., Stanley, M. L., Yu, H., & Crockett, M. J. (2024). *Motivated abstraction of personal transgressions in moral narratives*.

Currie, G. (2010). *Narratives and narrators: A philosophy of stories*. USA: Oxford University Press.

Dahlstrom, M. F. (2021). The narrative truth about scientific misinformation. *Proceedings of the National Academy of Sciences, 118*(15), Article e1914085117.

Davis, W. A. (2016). Implicature. *Irregular Negatives, Implicatures, and Idioms*, 51–84.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin, 129*(1), 74.

Dunbar, R. I. (2004). Gossip in evolutionary perspective. *Review of General Psychology, 8*(2), 100–110.

Ekman, P., Friesen, W. V., & O'sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology, 54*(3), 414.

Epley, N., & Dunning, D. (2000). Feeling" holier than thou": Are self-serving assessments produced by errors in self-or social prediction? *Journal of Personality and Social Psychology, 79*(6), 861.

Exline, J. J., Deshea, L., & Holeman, V. T. (2007). Is apology worth the risk? Predictors, outcomes, and ways to avoid regret. *Journal of Social and Clinical Psychology, 26*(4), 479–504.

Fraser, R. (2021). Narrative testimony. *Philosophical Studies, 178*, 4025–4052.

Fu, G., Xu, F., Cameron, C. A., Heyman, G., & Lee, K. (2007). Cross-cultural differences in children's choices, categorizations, and evaluations of truths and lies. *Developmental Psychology, 43*(2), 278.

Genette, G. (1980). *Narrative discourse: An essay in method. 1972. Trans. Jane E. Lewin*. Ithaca: Cornell UP.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79*(5), 701.

Green, M. C., & Donahue, J. K. (2011). Persistence of belief change in the face of deception: The effect of factual stories revealed to be false. *Media Psychology, 14*(3), 312–331.

Grice, H. P. (1957). Meaning. *The Philosophical Review, 66*(3), 377–388.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Griffin, D. J., & Bender, C. (2019). Culture and deception: The influence of language and societies on lying. *The Palgrave Handbook of Deceptive Communication*, 67–89.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development, 77*(3), 505–524.

Harris, P. L., Koenig, M. A., Corriveau, K. H., & Jaswal, V. K. (2018). Cognitive foundations of learning from testimony. *Annual Review of Psychology, 69*, 251–273.

Haugh, M. (2009). Intention (ality) and the conceptualization of communication in pragmatics. *Australian Journal of Linguistics, 29*(1), 91–113.

Hills, A. (2013). Moral testimony. *Philosophy Compass, 8*(6), 552–559.

Horn, L. R. (2010). Multiple negation in English and other languages. *The expression of negation*, 111–148.

Jaswal, V. K., & Neely, L. A. (2006). Adults don't always know best: Preschoolers use past reliability over age when learning new words. *Psychological Science, 17*(9), 757–758.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science, 21*(10), 1541–1547.

Kawamura, Y., Sasaki, S., & Kusumi, T. (2022). Cultural similarities and differences in lay theories of altruism: Replication of Carlson and Zaki (2018) in a Japanese sample. *Asian Journal of Social Psychology, 25*(3), 571–576.

Kelley, D. L., & Waldron, V. R. (2005). An investigation of forgiveness-seeking communication and relational outcomes. *Communication Quarterly, 53*(3), 339–358.

Kim, J., & Crockett, M. J. (2022). Narrating the" what" and" why" of our moral actions. In *, 44, no. 44. Proceedings of the Annual Meeting of the Cognitive Science Society*.

Knight, K. (2018). Transgressor communication after committing a hurtful relational event: Associations with attributed intent, blame, guilt, responsibility, and perceived forgiveness. *Western Journal of Communication, 82*(4), 475–492.

Koenig, M. A., Clément, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science, 15*(10), 694–698.

Lee, K., Xu, F., Fu, G., Cameron, C. A., & Chen, S. (2001). Taiwan and mainland Chinese and Canadian children's categorization and evaluation of lie-and truth-telling: A modesty effect. *British Journal of Developmental Psychology, 19*(4), 525–542.

Levine, S., Mikhail, J., & Leslie, A. M. (2018). Presumed innocent? How tacit assumptions of intentional structure shape moral judgment. *Journal of Experimental Psychology: General, 147*(11), 1728.

Li, L., & Tomasello, M. (2021). On the moral functions of language. *Social Cognition, 39*(1), 99–116.

Lin-Healy, F., & Small, D. A. (2013). Nice guys finish last and guys in last are nice: The clash between doing well and doing good. *Social Psychological and Personality Science, 4*(6), 692–698.

Loy, J. E., Rohde, H., & Corley, M. (2018). Cues to lying may be deceptive: Speaker and listener behaviour in an interactive game of deception. *Journal of Cognition, 1*(1).

Luyten, P., Sabbe, B., Blatt, S. J., Meganck, S., Jansen, B., De Grave, C., … Corveleyn, J. (2007). Dependency and self-criticism: Relationship with major depressive disorder, severity of depression, and clinical presentation. *Depression and Anxiety, 24*(8), 586–596.

MacGregor, J. C., & Holmes, J. G. (2011). Rain on my parade: Perceiving low self-esteem in close others hinders positive self-disclosure. *Social Psychological and Personality Science, 2*(5), 523–530.

Mahr, J. B., & Csibra, G. (2020). Witnessing, remembering, and testifying: Why the past is special for human beings. *Perspectives on Psychological Science, 15*(2), 428.

Mansfield, C. D., McLean, K. C., & Lilgendahl, J. P. (2010). Narrating traumas and transgressions: Links between narrative processing, wisdom, and well-being. *Narrative Inquiry, 20*(2), 246–273.

Mansfield, C. D., Pasupathi, M., & McLean, K. C. (2015). Is narrating growth in stories of personal transgressions associated with increased well-being, self-compassion, and forgiveness of others? *Journal of Research in Personality, 58*, 69–83.

Marsh, E. J., Meade, M. L., & Roediger, H. L., III (2003). Learning facts from fiction. *Journal of Memory and Language, 49*(4), 519–536.

Marsh, E. J., & Tversky, B. (2004). Spinning the stories of our lives. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 18*(5), 491–503.

Mascaro, O., Morin, O., & Sperber, D. (2017). Optimistic expectations about communication explain children's difficulties in hiding, lying, and mistrusting liars. *Journal of Child Language, 44*(5), 1041–1064.

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition, 112*(3), 367–380.

Mazzarella, D. (2015). Pragmatics and epistemic vigilance: The deployment of sophisticated interpretative strategies. *Croatian Journal of Philosophy, 15*(44), 183–199.

Mazzarella, D., & Pouscoulous, N. (2021). Pragmatics and epistemic vigilance: A developmental perspective. *Mind & Language, 36*(3), 355–376.

Mazzarella, D., Reinecke, R., Noveck, I., & Mercier, H. (2018). Saying, presupposing and implicating: How pragmatics modulates commitment. *Journal of Pragmatics, 133*, 15–27.

Mazzarella, D., & Vaccargiu, E. (2024). Communication: Inferring speaker intentions or perceiving the world? Insights from developmental research. *Journal of Pragmatics, 221*, 123–136.

Meibauer, J. (2014). *Lying at the semantics-pragmatics interface. 14*. Walter de Gruyter GmbH & Co KG.

Mercier, H. (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.

Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological science, 25*(3), 648–655.

Owens, T. J. (1993). Accentuate the positive-and the negative: Rethinking the use of self-esteem, self-deprecation, and self-confidence. *Social Psychology Quarterly*, 288–299.

Pasupathi, M., McLean, K. C., & Weeks, T. (2009). To tell or not to tell: Disclosure and the narrative self. *Journal of Personality, 77*(1), 89–124.

Pasupathi, M., & Wainryb, C. (2010). Developing moral agency through narrative. *Human Development, 53*(2), 55–80.

Paulhus, D. L. (1991). Measurement and control of response Bias. *Measures of Personality and Social Psychological Attitudes, 1*, 17–59.

Reboul, A. (2011). A relevance-theoretic account of the evolution of implicit communication. *Studies in Pragmatics, 13*(1), 1–19.

Reboul, A. (2017). Is implicit communication a way to escape epistemic vigilance. *Pragmatics at its Interfaces, 17*, 91.

Rimé, B., Finkenauer, C., Luminet, O., Zech, E., & Philippot, P. (1998). Social sharing of emotion: New evidence and new questions. *European Review of Social Psychology, 9*(1), 145–189.

Roberts, A. R., Levine, E. E., & Sezer, O. (2021). Hiding success. *Journal of Personality and Social Psychology, 120*(5), 1261.

Rui, J., & Stefanone, M. A. (2013). Strategic self-presentation online: A cross-cultural study. *Computers in Human Behavior, 29*(1), 110–118.

Schlenker, B. R. (2003). Self-presentation. *Handbook of Self and Identity, 2*, 542–570.

Schlenker, B. R., & Leary, M. R. (1982). Audiences' reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal of Experimental Social Psychology, 18*(1), 89–104.

Searle, J. R. (1979). Intentionality and the use of language. In *Meaning and use: Papers presented at the second Jerusalem philosophical encounter April 1976* (pp. 181–197). Dordrecht: Springer Netherlands.

Sedikides, C., Hoorens, V., & Dufner, M. (2015). *Self-enhancing self-presentation: Interpersonal, relational, and organizational implications*.

Shafto, P., Eaves, B., Navarro, D. J., & Perfors, A. (2012). Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental Science, 15*(3), 436–447.

Shaw, A., Montinari, N., Piovesan, M., Olson, K. R., Gino, F., & Norton, M. I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology: General, 143*(1), 363.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of self-defense: Self-affirmation theory. *Advances in experimental social psychology, 38*, 183–242.

Speer, S. A. (2019). Reconsidering self-deprecation as a communication practice. *British Journal of Social Psychology, 58*(4), 806–828.

Sperber, D. (2001). An evolutionary perspective on testimony and argumentation. *Philosophical Topics, 29*(1/2), 401–413.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language, 25*(4), 359–393.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.

Steinmetz, J., Sezer, O., & Sedikides, C. (2017). Impression mismanagement: People as inept self-presenters. *Social and Personality Psychology Compass, 11*(6), Article e12321.

Stellar, J. E., & Willer, R. (2018). Unethical and inept? The influence of moral information on perceptions of competence. *Journal of Personality and Social Psychology, 114*(2), 195.
Tice, D. M., Butler, J. L., Muraven, M. B., & Stillwell, A. M. (1995). When modesty prevails: Differential favorability of self-presentation to friends and strangers. *Journal of Personality and Social Psychology, 69*(6), 1120.
Vallauri, E. L., Laura, B., Cimmino, D., Federica, C., Claudia, C., & Giorgia, M. (2020). Implicit argumentation and persuasion a measuring model. In *Argumentation and meaning: Semantic and pragmatic reflexions* (pp. 95–123).
Van Damme, C., Hoorens, V., & Sedikides, C. (2016). Why self-enhancement provokes dislike: The hubris hypothesis and the aversiveness of explicit self-superiority claims. *Self and Identity, 15*(2), 173–190.
Vanderbilt, K. E., Liu, D., & Heyman, G. D. (2011). The development of distrust. *Child Development, 82*(5), 1372–1380.
Vonk, R. (1999). Impression formation and impression management: Motives, traits, and likeability inferred from self-promoting and self-deprecating behavior. *Social Cognition, 17*(4), 390–412.
Vrij, A. (2008). *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons.
Vullioud, C., Clément, F., Scott-Phillips, T., & Mercier, H. (2017). Confidence as an expression of commitment: Why misplaced expressions of confidence backfire. *Evolution and Human Behavior, 38*(1), 9–17.
Watling, D., & Banerjee, R. (2007). Children's understanding of modesty in front of peer and adult audiences. *Infant and Child Development: An International Journal of Research and Practice, 16*(3), 227–236.
Wiessner, P. W. (2014). Embers of society: Firelight talk among the Ju/'hoansi bushmen. *Proceedings of the National Academy of Sciences, 111*(39), 14027–14035.
Wolf, S. T., Cohen, T. R., Panter, A. T., & Insko, C. A. (2010). Shame proneness and guilt proneness: Toward the further understanding of reactions to public and private transgressions. *Self and Identity, 9*(4), 337–362.
Yerkovich, S. (1977). Gossiping as a way of speaking. *The Journal of Communication, 27*(1), 192–196.
Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind, 4*, 71–87.
Yus, F. (1999). Misunderstandings and explicit/implicit communication. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA), 9*(4), 487–517.
Zell, A. L., & Exline, J. J. (2010). How does it feel to be outperformed by a "good winner"? Prize sharing and self-deprecating as appeasement strategies. *Basic and Applied Social Psychology, 32*(1), 69–85.
Zhang, Y., Chen, X., Liu, Z., Zhang, Y., Jiang, T., You, X., & Luo, Y. (2021). Nice guys finish last? The effect of lay theories on prosocial actors' motivation and future benefits. *Social Psychological and Personality Science, 12*(1), 91–98.