COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e70010

ISSN: 1551-6709 online
DOI: 10.1111/cogs.70010

How Likely Is it that I Would Act the Same Way: Modeling Moral Judgment During Uncertainty

Paul C. Bogdan,[a,b] Sanda Dolcos,[a,b,c] Florin Dolcos[a,b,c]

[a] *Beckman Institute for Advanced Science and Technology, University of Illinois Urbana-Champaign*
[b] *Department of Psychology, University of Illinois Urbana-Champaign*
[c] *Neuroscience Program, University of Illinois Urbana-Champaign*

Received 1 June 2024; received in revised form 1 October 2024; accepted 17 October 2024

Abstract

Moral rules come with exceptions, and moral judgments come with uncertainty. For instance, stealing is wrong and generally punished. Yet, it could be the case that the thief is stealing food for their family. Such information about the thief's context could flip admonishment to praise. To varying degrees, this type of uncertainty regarding the context of another person's behavior is ever-present in moral judgment. Hence, we propose a model of how people evaluate others' behavior: We argue that individuals principally judge the righteousness of another person's behavior by assessing the likelihood that they would act the same way if they were in the person's shoes. That is, if you see another person steal, you will consider the contexts where you too would steal and assess the likelihood that any of these contexts are true, given the available information. This idea can be formalized as a Bayesian model that treats moral judgment as probabilistic reasoning. We tested this model across four studies ($N = 601$) involving either fictional moral vignettes or economic games. The studies yielded converging evidence showing that the proposed model better predicts moral judgment under uncertainty than traditional theories that emphasize social norms or perceived harm/utility. Overall, the present studies support a new model of moral judgment with the potential to unite research on social judgment, decision-making, and probabilistic reasoning. Beyond this specific model, the present studies also more generally speak to how individuals parse uncertainty by integrating across different possibilities.

Keywords: Moral judgment; Behavioral economics; Trust; Reciprocity; Bayesian inference

Correspondence should be sent to Paul C. Bogdan, Beckman Institute for Advanced Science & Technology, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801, USA. E-mail: pbogda2@illinois.edu

1. Introduction

Moral judgment is always clouded by some degree of uncertainty. It is simply impossible to perfectly understand another person's view of the world and how it motivated their decision. Nonetheless, humans readily blame others based on partial information, even though missing contextual details could flip virtually any act deemed immoral to instead be judged positively or vice versa. Clarifying how individuals grapple with this uncertainty is critical to understanding social interaction (FeldmanHall & Shenhav, 2019; Jara-Ettinger, 2019; Jara-Ettinger, Schulz, & Tenenbaum, 2020; Kim, Park, & Young, 2020). However, moral judgment theories usually ignore uncertainty. We instead developed and tested a model of moral judgment that places uncertainty in a central position (Kim et al., 2020; Kim, Mende-Siedlecki, Anzellotti, & Young, 2021). Taking inspiration from Bayesian models of Theory of Mind, we posit that individuals grapple with uncertainty during moral judgment by imagining different possible contextual explanations for another person's behavior and assessing the probability of each one being true. In turn, we argue that individuals judge the goodness of another person's action based on the likelihood that they would act the same way if they were in the other person's context. Below, we elaborate on this idea and describe how it was tested over four studies.

1.1. Moral judgment and inference about uncertainty

To investigate the moral domain, the present study leverages principles from earlier research on Bayesian models of Theory of Mind (Baker, Saxe, & Tenenbaum, 2011; Kim et al., 2021; Kim et al., 2020). Such models posit that, after observing another person's behavior, individuals use Bayesian inference to draw conclusions about the person and the person's context. Such models also posit that people's prior beliefs inform their interpretations of others' behavior, which is often a point of key interest (e.g., for understanding stereotyping). Probabilistic pathways linked to Theory of Mind bear importance to moral judgment. As an illustration, consider an individual who just observed someone flee a department store, stealing an unknown item. Suppose the observer optimistically believes that 90% of people would steal food if their family was hungry and believes only 1% would steal a luxury good for themselves out of simple greed. Given these prior beliefs, the optimistic observer would rationally infer that it is 90 times more likely that the thief is stealing food for their family than selfishly stealing a luxury good for themselves. Now, suppose instead that the thief was spotted by a pessimistic observer who reasons that the thief is most likely stealing a luxury good out of greed rather than food for their family. Of these two observers, the optimistic one would presumably judge the thief less harshly, meaning that moral judgment was guided by how the two observers parsed uncertainty. This thought experiment is intuitive, but there is currently limited experimental evidence demonstrating that this type of probabilistic inference indeed occurs during moral judgment. The present research investigates this topic and hypothesizes a specific computational mechanism that links participants' inferences about possible explanations to their judgments.

We specifically propose that individuals judge others' behavior based on the likelihood that they would act the same way if they were in the other person's position. This idea is motivated by research suggesting that the pursuit of behavioral similarity is a fundamental driver of social interaction and often guides judgments. For instance, reciprocity (treating others how they treated you) and the expectation of reciprocity (wanting others to treat you how you treated them) are seen as morally appropriate across the globe (Curry, Mullins, & Whitehouse, 2019) and emerge even among infants (Hamlin & Wynn, 2011). Beyond just reciprocity, viewing behavior similar to one's own, such as seeing another person mimic one's language or body movements, promotes positive impressions across a wide variety of contexts (Bocian, Baryla, Kulesza, Schnall, & Wojciszke, 2018; Fischer-Lokou, Martin, Guéguen, & Lamy, 2011; Kulesza et al., 2022; Kulesza, Dolinski, Huisman, & Majewski, 2014; Quiros, Kapcak, Hung, & Cabrera-Quiros, 2021). Studies have even shown that selfish individuals prefer people who are also selfish more than others who are generous (Herrmann, Thöni, & Gächter, 2008; Irwin & Horne, 2013; Monin, Sawyer, & Marquez, 2008). These diverse results on the effects of similarity on social judgment motivate our hypothesis.

Earlier theories of moral judgment have been premised on other foundations. For example, several previous theories have focused on social norms. These theories posit that individuals evaluate others' behavior by reflecting on descriptive social norms and, in turn, individuals tend to negatively judge behavior that is atypical (Goldring & Heiphetz, 2020; Gollwitzer, Martel, Bargh, & Chang, 2020; Lindström, Jangard, Selbing, & Olsson, 2018). Previous experiments have indeed identified circumstances where typicality predicts the intensity of moral judgments (Vavra, Chang, & Sanfey, 2018; Xiang, Lohrenz, & Montague, 2013). However, there are issues with this explanation: Individuals sometimes prefer norm violators if they themselves also previously violated those norms (Monin et al., 2008). Intuitively, one can also imagine cases where believing that some behavior is widespread would actually lead to harsher judgment. For instance, a parent may suspect that drug use is rampant, but this presumably would not lead them to condone their children consuming drugs. Rather, the parent may even become stricter about uncertain suspicious behavior, as they would be predisposed to assume their child is doing drugs. This outcome would instead speak to the proposed self-likelihood model and its focus on Bayesian reasoning. Regardless, the norm-violation view is of special interest to the present research, as it predicts that social judgment is proportional to people's perception of the likelihood most people would perform a given action. This parallels the proposed self-likelihood idea, and hence the norm violation view would be the primary point of comparison.

1.2. Present research

Four studies were conducted to investigate moral judgment under uncertainty. Each study had one part where participants were told about another person's behavior in an "uncertain context." Based on this uncertain information, participants were asked to evaluate the person's behavior, such as by judging their trustworthiness. In each study, participants also were told of "possible contexts" that could be behind the other person's behavior. For these contexts, participants were asked about how they would act and how they predict most people would act. For instance, in the thief example above, one "possible context" is that the thief's family needed food. Based on participant's responses to the different parts of the task, we tested the proposed *self-likelihood model*. As described, the model posits that participants evaluate others' behavior during uncertainty by drawing inferences about possible contexts to assess the likelihood that they would act the same way. In all four studies, we modeled this perceived likelihood, measured whether it significantly predicted participants' moral judgments, and compared the model's accuracy to the accuracy of alternative views (e.g., the norm-violation view).

Every study followed this same structure in design and analysis, but the studies each had some unique features: Study 1 used moral vignettes, and this study aims to provide initial evidence demonstrating the viability of the self-likelihood model for tackling different aspects of moral judgment (e.g., empathy/respect/etc.). Study 2, used economic games, and it served to test more precise aspects of the model, such as examining whether participants' judgments indeed reflect probabilistic inference about possible contexts. Finally, Studies 3A and 3B probe the extent of the model's generalizability, investigating different dependent variables and drawing comparisons to alternative theories beyond the norm-violation view.

2. Study 1

Study 1 compared the proposed self-likelihood model to the norm-violation view using a moral vignette design where participants responded to two questionnaires. First, participants completed an "Evaluation questionnaire," wherein they judged a character's behavior in an uncertain context. Second, participants completed a "Decision & Prediction questionnaire," wherein they read about more specific *possible contexts*, stated how they would act themselves, and predicted how most people would behave. Each uncertainty-filled vignette from the first questionnaire mapped onto 10 possible contexts in the second questionnaire. This design permits testing the self-likelihood model by attempting to link participants' evaluations to their decisions across the 10 corresponding possible contexts. In addition, the design permits testing the norm-violation model because the task also probes participants' predictions about typical behavior. The design does not yet allow investigating Bayesian inferences about each possible context's perceived likelihood, but later studies would investigate this finer point.

2.1. Methods

2.1.1. Participants

One hundred sixty students were recruited from the local university and completed this study online (M_{age} = 19.1 [18–23]; 61% Female, 38% Male, 1% Nonbinary; 29% Asian, 23% Hispanic, 11% Black, 31% White, 6% Multiracial/no-response). Power analyses ($\alpha = .05$, 80% power) using data from a preliminary version of the study revealed that 65 participants would be sufficient to detect significant fixed effects linked to the hypothesized models (see below); power calculations were done using the *simr* package (Green &

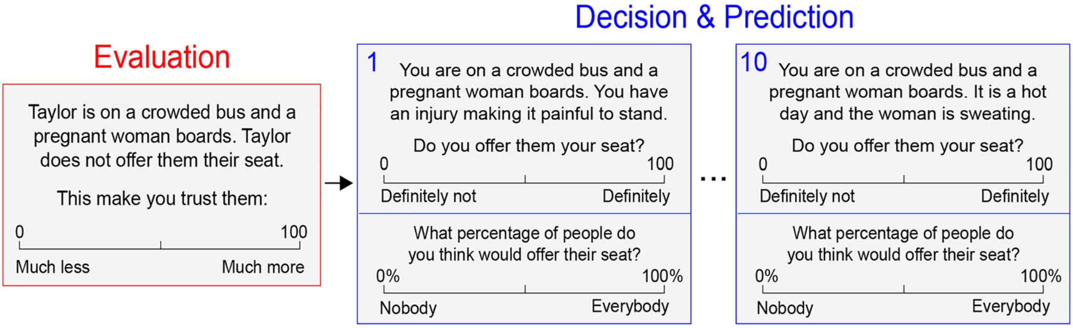
Evaluation
Taylor is on a crowded bus and a pregnant woman boards. Taylor does not offer them their seat.
This make you trust them:
0
100
Much less
Much more
Decision & Prediction
1
You are on a crowded bus and a pregnant woman boards. You have an injury making it painful to stand.
Do you offer them your seat?
0
100
Definitely not
Definitely
What percentage of people do you think would offer their seat?
0%
100%
Nobody
Everybody
...
10
You are on a crowded bus and a pregnant woman boards. It is a hot day and the woman is sweating.
Do you offer them your seat?
0
100
Definitely not
Definitely
What percentage of people do you think would offer their seat?
0%
100%
Nobody
Everybody

Fig. 1. Diagram of the Study 1 questionnaires. The left panel shows a simplified version of a vignette included in the Evaluation questionnaire. Each Evaluation vignette mapped onto 10 possible context questions in the Decision & Prediction questionnaire (two shown on the right). Note that the decision and prediction questions specifically referred to the specific context described in the blue boxes ("You have an injury…") and participants were never asked about the vague context described in the red box.

MacLeod, 2016). Data collection went beyond just 65 participants, continuing until the end of our university's credit pool period to increase statistical power. All participants provided informed consent under a protocol approved by the Institutional Review Board and received course credit for participation.[1] The data and R code needed to reproduce the results for every study in this report are available in a public repository (https://osf.io/h5mea).

2.1.2. *Task design*

Participants an Evaluation questionnaire and a Decision & Prediction questionnaire. The two questionnaires are provided in the aforementioned repository (https://osf.io/h5mea). The questions covered issues related to empathy (Fig. 1), authority/loyalty/empathy (helping a relative), authority/fairness (behaving respectfully in class), and fairness (returning inadvertently stolen goods).

For the Evaluation questionnaire, participants read four short realistic vignettes where a character acted inappropriately. Few details were given about the context of the character's behavior, making it unclear whether the behavior is mildly inappropriate or highly inappropriate (see Fig. 1 for an example). Participants were asked to imagine that the character is a peer and to report, using a continuous scale, how learning about this inappropriate behavior would change the extent to which they trusted the character (0 = "*Much less*," 50 = "*No change*," 100 = "*Much more*"). Trust was used as the dependent variable because it applies to both positive and negative judgments—unlike, for instance, asking about "moral wrongness," which can only measure the extent of negativity (for discussion on specific evaluation measures, see Malle, 2021).

After participants finished the Evaluation questionnaire, they began the Decision & Prediction questionnaire. Each vignette from the first phase was used to generate 10 possible contexts, which were more detailed (see Fig. 1 for two examples). For each possible context, participants were asked to report whether they themselves would behave in an appropriate manner (0 = "*Definitely not*"; 100 = "*Definitely would*") and to report what percentage of other people they think would perform the appropriate action in that context (0% = "*Nobody*"; 100% = "*Everybody*"). Participants completed the full Evaluation questionnaire before beginning the Decision & Prediction questionnaire to avoid directly encouraging participants to consider specific possible contextual details when processing the uncertain vignette.

2.1.3. Defining the models

For each evaluation, two variables (*Self* & *Norm*) were calculated based on the participant's responses for the corresponding 10 possible contexts. *Self* represents the proposed model. One version of *Self* was computed based on averaging participants' decision-making responses across the 10 possible contexts (reverse coding of the 0-to-100 decision scale asking about appropriate behavior). In addition, a separate version of *Self* was tested which treated decision-making as binary and focused on the proportion of possible contexts where the participant said that they would likely behave inappropriately (*No* = under 50, *Yes* = over 50; responses of exactly 50 were discarded). *Norm* represents the norm-based view and was similarly defined while either treating the measure as continuous or binary.

Notably, *Self* and *Norm* are themselves correlated (continuous forms: $r = .62$ [.56, .66]), which may reflect conformity (*Norm* → *Self*) or a false consensus effect (*Self* → *Norm*). Such biases are an aspect of social cognition, which are considered in the General Discussion. For research focusing on this conformity effect and its mechanisms within a similar evaluation and decision design, see Bogdan et al. (2022). In addition, see Bogdan et al. (2023) for research similar research predicting social judgments that likewise compared the effects of comparisons to oneself to comparisons with normative beliefs, but instead modeled normative beliefs relative to prior observations rather than stated perceptions.

2.1.4. Regressions

To assess whether each variable significantly predicted trust judgment, separate multilevel regression was fit, formatted here as R-style *lme4* equations,

$$Trust \sim 1 + Self + (1 + Self|Participant) + (1 + Self|Vignette)$$

and

$$Trust \sim 1 + Norm + (1 + Norm|Participant) + (1 + Norm|Vignette)\,.$$

For each regression, each participant contributed four data points (one for each of the four evaluations). The regressions included random slopes for their associated predictor because this is necessary for evaluating the significance of predictor fixed effects while not inflating the Type I error rate (Barr, Levy, Scheepers, & Tily, 2013; Meteyard & Davies, 2020). Additionally, the use of random slopes ensures that emerging associations generalize across the four vignettes. For each regression, the significance of its fixed effects was measured using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017).

Self and *Norm* were also compared in terms of their model fits. For these analyses, the regressions were tested using the same random intercepts as above but now without random slopes,

$$\mathit{Trust} \sim 1 + \mathit{Self} + (1|\mathit{Participant}) + (1|\mathit{Vignette})$$

and

$$\mathit{Trust} \sim 1 + \mathit{Norm} + (1|\mathit{Participant}) + (1|\mathit{Vignette})\,.$$

Random slopes, although necessary for testing the significance of fixed effects, would limit clear interpretation of model fit. For instance, if random slopes were used, then fit could be increased by a participant who shows a linear but negative relationship between *Trust* and *Self*, which conflicts with the directionality posited by the model (positive judgments are associated with a *greater* perceived likelihood of acting the same way). Model fit was measured as log-likelihood; for reference, a log-likelihood difference greater than 2.3 is typically considered strong, and a difference greater than 4.6 is typically considered decisive (Kass & Raftery, 1995).

2.2. *Results*

The proposed self-likelihood model, which focuses on possible contexts and participants' own decision-making, effectively predicted participants' evaluations of inappropriate behavior (Fig. 2). Specifically, the regression coefficients showed significant effects of self-likelihood on trust evaluations regardless of whether the analyses treat decision-making as continuous (β [standardized] = .34 [.25, .43], $p < .001$) or binary ($\beta = .27$ [.16, .39], $p = .007$). On the other hand, participants' perceived likelihood that a typical person would behave the same way did not significantly predict trust evaluations (continuous: $\beta = .11$ [.00, .22], $p = .07$; binary: $\beta = .07$ [−.05, .20], $p = .26$). Furthermore, direct comparison of the models showed that the regressions using the self-likelihood predictor yielded stronger model fits (Fig. 2, right side).

2.3. *Discussion*

These findings provide initial evidence that participants' evaluations of behavior in uncertain contexts are linked to their own (reported) decision-making in possible contexts, which supports the proposed self-likelihood model. The use of moral vignettes for this study is partly a strength, given that they can probe a variety of different moral topics and situations that are difficult to create in a laboratory. However, moral vignettes also carry limitations. Most obviously, they may fail to capture genuine behavior (Bostyn, Sevenhant, & Roets, 2018; FeldmanHall et al., 2012). Additionally, and specifically relevant to the present research, the vignette design does not allow us to test for Bayesian inference about possible contexts' perceived likelihoods. This is because we cannot quantify or control participants' priors about the possible contexts. We addressed this in Study 2, which overcame these limitations by using an economic game where participants' priors can be established as uniform across all possible contexts.

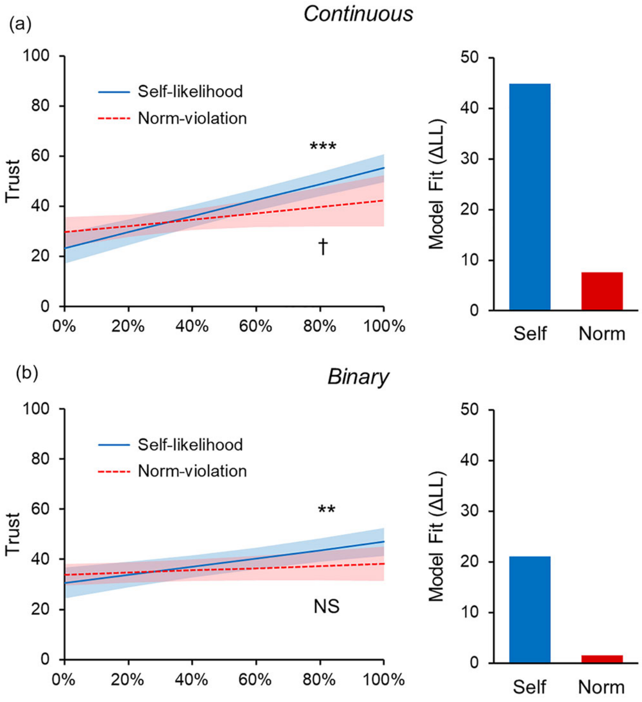

Fig. 2. Self-likelihood better predicted judgment of character. (A) The line and bar graphs at the top correspond to the analysis done using the continuous versions of the *Self* and *Norm* variables. On the left, the *self-likelihood* line indicates that participants judged characters more positively if the participants were likely to behave the same way. Shaded areas represent 95% confidence intervals. On the right, the bars show that self-likelihood better predicted participants' judgments, as evidenced by its regression yielding a greater model fit, which is reported as the log-likelihood difference (ΔLL) relative to a plain model with only random intercepts. (B) The line and bar graphs on the bottom correspond to the same analyses but were done using the binary versions of *Self* and *Norm*. **, $p < .01$; ***, $p < .001$.

3. Study 2

Study 2 further probed the self-likelihood model. Like in the first study, analyses examined possible contexts and whether participants' judgments are more tied to their decision-making or predictions about other people. Study 2 additionally included analyses assessing whether participants draw inferences about possible contexts' likelihoods when evaluating whether they would behave the same way as another person. The analyses specifically examined whether participants' decision-making in possible contexts with a high perceived likelihood had a greater bearing on their evaluations than decision-making in possible contexts with a low perceived likelihood. In other words, if *Possible Context 1* is perceived to be a more likely explanation for another person's behavior than *Possible Context 2*, then participants' decision-making in the former should have a greater impact on their evaluation. Such results would provide evidence for the self-likelihood model, and more broadly,

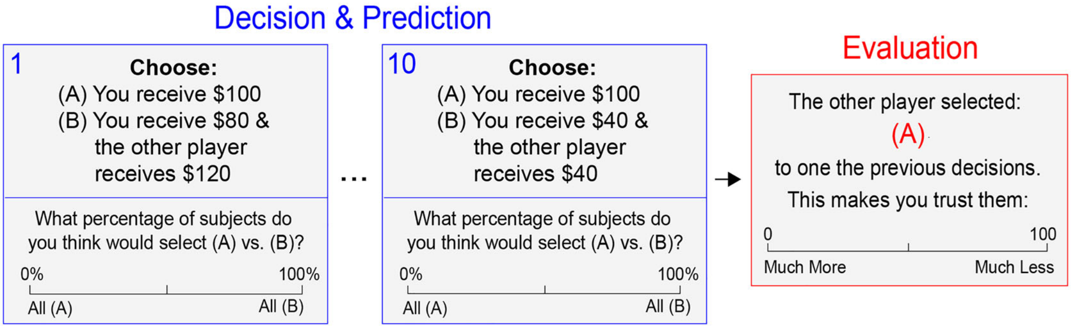
Decision & Prediction
1
Choose:
(A) You receive $100
(B) You receive $80 & the other player receives $120
What percentage of subjects do you think would select (A) vs. (B)?
0%
100%
All (A)
All (B)
...
10
Choose:
(A) You receive $100
(B) You receive $40 & the other player receives $40
What percentage of subjects do you think would select (A) vs. (B)?
0%
100%
All (A)
All (B)
Evaluation
The other player selected:
(A)
to one the previous decisions.
This makes you trust them:
0
100
Much More
Much Less

Fig. 3. Diagram of each Study 2 round. The game was organized as rounds. Each round had a Decision & Prediction phase followed by an Evaluation phase. Unlike in Study 1, participants alternated between the two phases—that is, completed the first set of Decision & Prediction questions, then the first evaluation, then the second set of Decision & Prediction questions, then the second evaluation, and so on. Note, participants were told that the other player completed the same 10 decisions, meaning that they knew overall which options were presented to them, but participants were not told which decision was selected.

support the idea that individuals parse uncertainty via Bayesian inference about possible contexts.[2]

3.1. Methods

3.1.1. Participants

Three hundred thirty-four students (M_{age} = 19.1 [18–24]; 62% Female, 36% Male, 2% Nonbinary; 41% Asian, 6% Black, 12% Hispanic, 38% White, 3% Multiracial/no-response) were recruited from the local university and completed this online study. Power analyses ($\alpha = .05$, 80% power) using data from a preliminary version of the study revealed that 81 participants would be sufficient to detect significant fixed effects linked to the hypothesized models. However, data collection continued until the end of our university's credit pool period to increase statistical power. All participants provided informed consent under a protocol approved by the Institutional Review Board and received course credit for participation.

3.1.2. Task design

Participants completed a two-player economic game, wherein they made monetary decisions, predicted their partner's decisions, and evaluated their partner's decisions based on uncertain information. Although we were not allowed to have participants play the game for real money, participants were told that their earnings would determine how much time they would need to complete a boring task after the experiment. Time-saving incentives have been shown to elicit similar behavior as monetary compensation (Jouxtel, 2019; Noussair & Stoop, 2015).

The game was organized into 12 rounds (Fig. 3). Each round consisted of a Decision & Prediction phase and then an Evaluation phase. For the first phase, participants made binary choices in 10 contexts. Every context followed the same structure: Option A always awarded the participant $100 and their partner $0. Option B always led to both players receiving some amount of money, but the exact amounts varied between contexts. Along with making a decision, participants were asked to predict what percentage of other participants would select Option A or B for a given context. Participants were told that this prediction would not impact their earnings but were asked to still take it seriously. For each round, participants were told that their response to 1 of the 10 questions would be randomly selected, with equal probability, to impact their earnings and the earnings of a future participant.[3]

Immediately after each round's Decision & Prediction phase, participants began the round's Evaluation Phase. Participants were told that another player completed the same 10 decisions and one of the player's decisions was randomly selected to impact the participant. Participants were told whether this player chose Option A or B but were not told which decision the response was directed to. Because each decision had a similar structure, where Option A was always a selfish choice, participants could immediately recognize whether the player's choice was selfish or generous. However, the extent of selfishness or generosity was uncertain. For example, if the question offered, "*(B) You receive $80 & the other player receives $120*," then selecting *Option A* would be highly selfish, as it implies a refusal to even give up $20 so that the other player could receive $120. By contrast, if the question offered, "*(B) You receive $40 & the other player receives $40*," then selecting Option A would not be as selfish. Based on this uncertain information about the context of the other player's decision, participants were asked to report the player's perceived trustworthiness using a 0–100 scale. Participants were told that this evaluation would not influence earnings but asked to still take it seriously. The exact dollar amounts, which could objectively quantify the degree of selfishness, were not analyzed here, although this would be done in Study 3.

Participants were told that they were playing with other people who changed between rounds, but in reality, they interacted with computers. Every participant completed the same 12 rounds. The 12 rounds were organized such that there were six configurations of 10 contexts, and after half of them, participants evaluated a player who purportedly chose Option A, and in the other half, they evaluated a player who purportedly chose Option B. Note that although judgment of immoral acts is typically the primary focus of social judgment research, participants would find it strange if they only encountered others behaving selfishly, and hence the design needed to include evaluations of both selfish (Option A) and generous (Option B) choices.

3.1.3. Defining the models

Analysis followed a similar procedure as in the first study, defining the norm violation (*Norm*) and self-likelihood (*Self*) models as variables and using them to predict trust evaluations in multilevel regressions, which are provided further below. As in Study 1, *Norm* and *Self* were positively correlated, and this applied to both selfish A-choice actions ($r = .63$ [.60, .66]) and generous B-choice actions ($r = .62$ [.59, .65]).

Along with comparing *Norm* versus *Self*, further tests were done to assess whether Bayesian inference occurred while participants assessed their likelihood of having acted the same way in their partner's context. First, $Self_{weighted}$ was defined, which calculated the overall likelihood of acting the same way while integrating across the possible contexts and weighing each one by its perceived likelihood,

$$Self_{weighted}\,(Act) = \sum_{1}^{10} \delta_{self}\,(action|context_i)\, p_{norm}(context_i|action),$$

which, given Baye's rule, is equivalent to

$$Self_{weighted}\,(Act) = \frac{1}{Z} \sum_{1}^{10} \delta_{self}\,(action|context_i)\, p_{norm}\,(action|context_i)\,.$$

$\delta_{self}(action|context_i)$ is 1 if a participant would perform the evaluated action in possible context i and 0 otherwise. $p_{norm}(action|context_i)$ represents a participant's prediction about the likelihood that a typical person would perform the action being evaluated in context i. Z represents the sum of $p_{norm}(action|context_i)$ across the 10 possible contexts and normalizes the equation to [0, 1]. Note that $p_{norm}(context_i|action)$ is equal to $\frac{1}{Z}$ $p_{norm}(action|context_i)$ per Baye's rule because participants were told that the 10 possible contexts were all equally likely to be selected, establishing a uniform prior regarding $p(context_i)$.[4]

The above equation assumes that participants accurately report their beliefs about the other person's behavior. However, it is known that there are biases in these types of probabilistic beliefs (see Prospect Theory, Tversky & Kahneman, 1992). This aspect is more relevant to $Self_{weighed}$ than to *Norm*, as *Norm* is inserted directly into a regression. Therefore, a uniform shift downward (e.g., of 10%) or differences in the standard deviation of reported probabilities (e.g., using the full 0–100% range vs. just 40–60%) will have no impact on the statistical significance or model fit for *Norm*. Yet, these types of biases introduce noise to $Self_{weighted}$.

Hence, an additional analysis was done, which defined two versions of *Self*: $Self_{low}$ considers the five possible contexts with the lowest perceived likelihood and in what proportion of these did participants report that they would behave the same way as the action being evaluated. These amount to the five contexts with the lowest predictions that the other player would perform the evaluated action, as $p(context \mid context) \propto p(action \mid context)$. Conversely, $Self_{high}$ considers the five possible contexts with the highest perceived likelihood. $Self_{low}$ and $Self_{high}$ were compared in terms of which better predicts trust evaluations. If participants are indeed performing Bayesian inference to assess the likelihood of each possible context, then their own behavior in the high-likelihood ones ($Self_{high}$) should be more predictive than their behavior in the low-likelihood possible contexts ($Self_{low}$). Here, the underlying assumption is not that participants can report their probabilistic beliefs accurately in an absolute sense but simply that the relationship between reported beliefs and "true" beliefs is monotonic (Tversky & Kahneman, 1992).

3.1.4. Regressions

Analysis paralleled the procedures used for Study 1. We submitted the variables (*Norm, Self*, $Self_{weighted}$, $Self_{low}$, & $Self_{high}$) to multilevel regressions. As before, we included random slopes when assessing the significance of fixed effects and excluded random slopes when comparing model fits. For instance, the regression with random slopes for *Self* was

$$\mathit{Trust} \sim 1 + \mathit{Self} + (1 + \mathit{Self}|\mathit{Participant}) + (1 + \mathit{Self}\,|\mathit{Round}).$$

Note that *Round* was defined to represent the twelve different round configurations, meaning that it controlled whether the participants evaluated a selfish (Option A) or generous (Option B) decision.

3.1.5. Exploratory test

Exploratory analyses considered the effects of a *Self* x *Norm* interaction, which can represent a participant perceiving similarity even when they overall behave in a non-normative manner across the most possible contexts. The meaningfulness of this interaction, how it differs from $\mathit{Self}_{weighted}$, and the associated results are reported in Supplementary Materials 1. Although effects are weak and mostly beyond the scope of our hypotheses, the findings suggest non-normative similarity has a particularly strong effect on trust, and this interaction between normative beliefs and decision-making is orthogonal to the main focus on Bayesian inference.

3.2. Results

The economic game data yielded further support for the self-likelihood model (Fig. 4). Examining the regression coefficients showed that trust was significantly predicted by both the self-likelihood model ($\beta = .15$ [.10, .20], $p < .001$; Fig. 4A) and the norm violation model ($\beta = .12$ [.07, .17], $p < .001$; Fig. 4B). However, a comparison of model fit revealed that the self-likelihood model better predicted participants' social judgments than the norm violation model (LL difference = 29.4, Fig. 4C).

Further comparisons dissected the self-likelihood model, testing specifically for whether participants used Bayesian inference to discern the likelihood of each possible context being the one responsible for their partner's decision. First, testing a version of the self-likelihood model that weighed each self-decision based on the perceived likelihood of each possible context yielded a significant association with trust ($\beta = .13$ [.07, .18], $p < .001$). The model's fit (ΔLL = 36.6) outperformed the norm-violation model (LL difference = 18.3; Fig. 4D). However, the fit does not surpass the original simpler self-likelihood model's fit (LL difference = -11.1). This drop in fit may be because participants cannot perfectly report their true beliefs about another person's behavior, which introduces noise. Hence, a more sensitive analysis of Bayesian inference was done which assumes that participants' predictions are at least ordinally correct. These tests yielded results consistent with the hypotheses: Participants' decision-making in possible contexts modeled as having a high perceived likelihood (Self_{high}) better predicted their judgments than their decision-making in possible contexts modeled as having a low perceived likelihood (Self_{low}; LL difference = 16.9, Fig. 4E). These results suggest that participants indeed use Bayesian inference to assess the likelihood of possible contexts when judging whether would have behaved the same way as the other person.

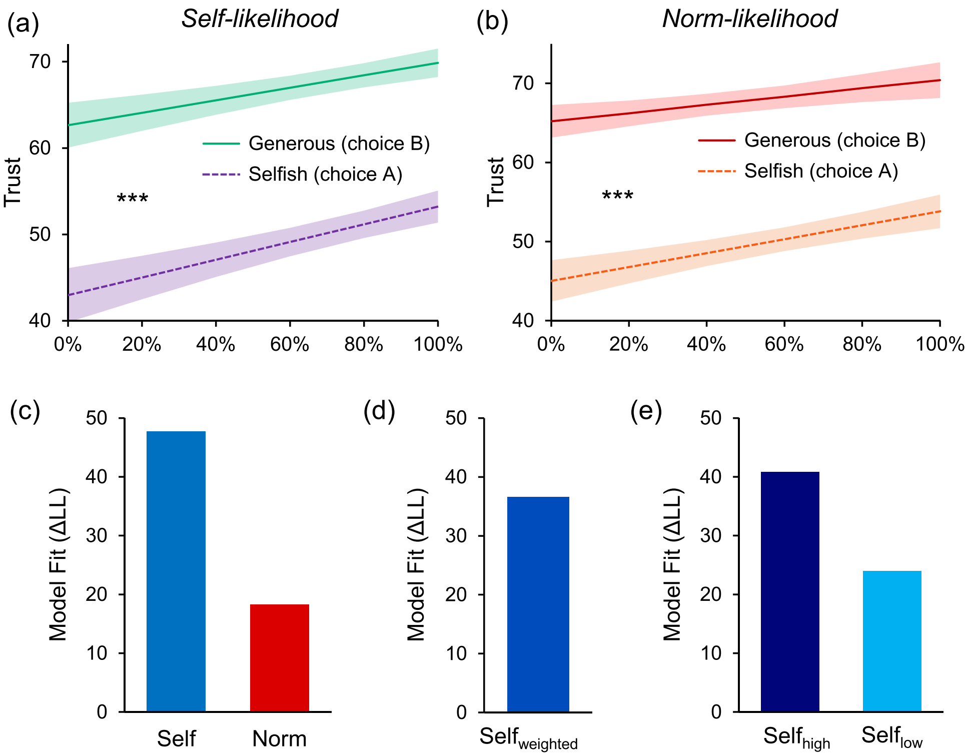

Fig. 4. Self-likelihood better predicted trust judgments. (A) Self-likelihood significantly predicted trust judgments when evaluating either other players' generous (choice B) or selfish (choice A) decisions. Shaded areas represent 95% confidence intervals. (B) Norm-likelihood also likewise predicted trust judgments. (C) Although both significant, self-likelihood (*Self*) predicted judgments better than perceived norm-violation (*Norm*), as evidenced by the former's regression yielding a greater model fit, which is reported as the log-likelihood difference (ΔLL) relative to a plain model with only random intercepts. (D) The self-likelihood variable, which weighs behavioral similarity based on the perceived likelihood of each possible context, effectively predicted trust albeit less so than *Self*. (E) Judgment is better predicted by self-likelihood modeled based on the five possible contexts with above-median perceived likelihoods ($Self_{high}$) relative self-likelihood modeled based on the contexts with below-median perceived likelihoods ($Self_{low}$). ***, $p < .001$.

3.3. Discussion

The findings support our proposed model that moral judgment involves participants reflecting on the likelihood that they too would perform the act being evaluated, and participants use Bayesian inference to assess this likelihood. As mentioned in the Methods, because the analyses simply compared the effects of high perceived likelihood versus low perceived likelihood possible contexts, this is a way to test for Bayesian inference while avoiding issues related to whether participants can genuinely express precise probabilistic beliefs. Interestingly, the norm violation predictor was also found to significantly predict trust evaluations. Although it was a weaker predictor than self-likelihood, this nonetheless speaks to prior studies on the role of norm violation in social judgment (Kawamura & Kusumi, 2020;

Walker, Turpin, Fugelsang, & Białek, 2021; Xiang et al., 2013). To further investigate and compare the proposed model to alternative perspectives, the upcoming studies adjusted the design of the Study 2 experiment to allow testing of two further alternative perspectives. In addition, because Studies 1 and 2 focused on solely evaluations of trust, it remains important to assess whether the effects here generalize across dependent variables, which the upcoming studies also investigated by probing fairness.

4. Studies 3A and 3B

The final two studies used a modified version of the Study 2 design and aimed to test alternative perspectives beyond just the norm-violation view. The changes specifically allow modeling *expectation-violation* theories, which argue that harsher judgments will occur when another person's behavior is how participants predict they will behave given their past decisions (Sytsma, Livengood, & Rose, 2012). In addition, the adjusted design permits modeling *consequentialist* theories, which argue that moral judgments will be harsher when another person's decision leads to greater harm (larger loss in earnings) for the participant (Cohen & Ahn, 2016; Engelmann & Waldmann, 2022; Schein & Gray, 2018). Along with comparing the self-likelihood model to a wider range of other theories, to evaluate the generalizability of our findings across different aspects of social evaluation, the study was conducted using two different dependent variables. Specifically, Study 3A examined how well the different models predicted trust, while Study 3B examined how well the models predicted fairness. Fairness differs from trustworthiness, in that it is a moral judgment directed to the action, not of the person behind it. This adds to the generalizability.

4.1. Methods

4.1.1. Participants

For Study 3A, 62 students were recruited from the local university for this online task. Data from two participants were excluded, one because they never replied to the decision screens and the other because they exited near the task's halfway point. This yielded a final sample of 60 participants (63% Female, 37% Male; $M_{age} = 19.6$ [18–22]; 27% Asian, 7% Black, 14% Hispanic, 52% White, 2% Multiracial/no response; does not sum to 100% due to rounding). For Study 3B, 45 new participants were recruited for this online task. Data from two participants were excluded, one for never responding to any evaluation screen and the other for failing to respond to every screen in nearly half of the trials. This yielded a final sample of 43 participants (56% Female, 44% Male; $M_{age} = 20.0$ [18–23]; 27% Asian, 18% Black, 12% Hispanic, 42% White).

These sample sizes are, notably, much smaller than those used in Studies 1 and 2, but in exchange, the studies here used a much larger number of trials (see below). Power analyses using pilot data generated with a preliminary version of the design ($N = 40$) showed that 12 participants would be sufficient to identify decisive differences in model fit (80% power, Δ log-likelihood > 4.6) between the self-likelihood model and the next best fitting model. Data collection went beyond this, continuing until the end of our university's credit pool period, to increase statistical power. All participants provided informed consent under a protocol approved by the Institutional Review Board and received course participation credit.

4.1.2. Task design

The Study 2 design was modified as follows: (a) Each round was changed so there were now just two possible contexts rather than 10. This simplified structure allowed participants to more easily and precisely see the impact of their partner's choice on their earnings. In turn, we were able to effectively test a *consequentialist model*, which posits that participants' judgments depend on the expected outcomes of the other player's choice. (b) Because each round only includes two possible contexts, 96 rounds could be included in the task, rather than just the 12 rounds used for Study 2. (c) Participants were told that they would repeatedly interact with the same partner and that they were playing with them in real-time. Participants were told that they would only change partners every 16 rounds. (d) Because the task now used a repeated-play design, participants could form impressions of the player. Accordingly, the prediction phase asked participants to predict how their partner specifically would behave. In turn, we could test an *expectation-violation* model, which posits that participants' judgments hinge on whether the other players' behavior was unexpected. This contrasts with the first and second studies, which asked participants to predict how most people would behave and tested a norm-violation model. (e) Finally, for Study 3A, the evaluation phase asked participants about perceived trustworthiness, mirroring Study 2, but for Study 3B, the evaluation phase now asked participants to judge whether their partner's choice was fair (0 = "*Very unfair*," 50 = "*Both fair and unfair*," 100 = "*Very fair*"). The evaluation question was the only difference between Studies 3A and 3B.

4.1.3. Defining the models

The proposed self-likelihood variable was calculated identically as in Studies 1 and 2, albeit now while only integrating across two possible contexts. As mentioned, two alternative models were also tested. First, an *Expectation-violation* variable was calculated identically to the norm-violation variable from Studies 1 and 2, but now while using participants' predictions about their specific partner rather than about their predictions about most people. Second, an *Outcome* (consequentialist) variable was calculated, which represents participants' economic gains/losses due to the other player's choice. These gains/losses are calculated relative to the choice that the other player did not select. For instance, consider a possible context where the participant would earn \$40 if the other player chose B. As always, the participant earns \$0 if the other player chooses A. Hence, from the participants' perspective, observing the other player Choose B constitutes a +40 gain, whereas if they choose A, this would be a −40 loss. We modeled the consequentialist value as the gains/losses averaged across each round's two possible contexts.

4.1.4. Regressions

The three models' variables were submitted to separate multilevel regressions predicting participants' Trust or Fairness responses. The regressions parallelled the structure of those used in Studies 1 and 2. They included random slopes for evaluating the significance of fixed effects and did not include random slopes for comparing model fits between regressions.

4.1.5. Exploratory and confirmatory analyses

The present task design lends itself to analyses of inequity and its role in social judgment. Supplementary Materials 2 provides the results of such analyses, which specifically use the Fehr−Schmidt (Fehr & Schmidt, 1999) model to estimate participants' inequity aversion along with participants' trial-by-trial perceptions of the other player's inequity aversion. Overall, the new results provide a further demonstration of how Bayesian inferences about self-other similarity guide social judgments. The results additionally show how the conclusions below are reproduced when Fehr−Schmidt model variables are included as covariates.

4.2. Results

The studies on trust and fairness generated converging results supporting the proposed self-likelihood model. First, examining the regressions' fixed effects showed that participants' likelihood of behaving the same way as their partner significantly predicted their changes in trust ($\beta = .18$ [.13, .23], $p < .001$) and their fairness judgments ($\beta = .27$ [.21, .33], $p < .001$) (Fig. 5 top & middle). Next, examining model fit showed that the self-likelihood model was more predictive than the expectation-violation view or the consequentialist model (Fig. 5 bottom). This was the case regardless of whether participants were evaluating trust or fairness.

4.3. Discussion

Studies 3A and 3B both provide further evidence favoring the proposed self-likelihood model, showing it better predicts moral judgment under uncertainty than expectation-violation or consequentialist theories. Additionally, Studies 3A and 3B show that our conclusions generalize beyond just judgments of another person's character (e.g., trust), but also to judgments of specific actions (e.g., fairness).

5. General discussion

Our research investigated how individuals judge the morality of another person's behavior when the context of the behavior is uncertain. We propose that to parse this uncertainty, individuals use Bayesian inference to assess the likelihood of various *possible contexts*. In turn, individuals judge the morality of the other person's behavior by integrating across these possible contexts and assessing the likelihood that they would act the same way. Study 1 modeled this idea and found evidence for it by examining participants' responses to moral vignettes. Study 2 found further support for the model and specifically demonstrated the role of inference about possible contexts in moral judgment. Finally, Studies 3A and 3B added yet more evidence by including new comparisons to other models and testing new dependent variables. Altogether, these studies showed that the proposed model is more predictive than theories focusing on norm violations, expectation violations, or consequences, and the

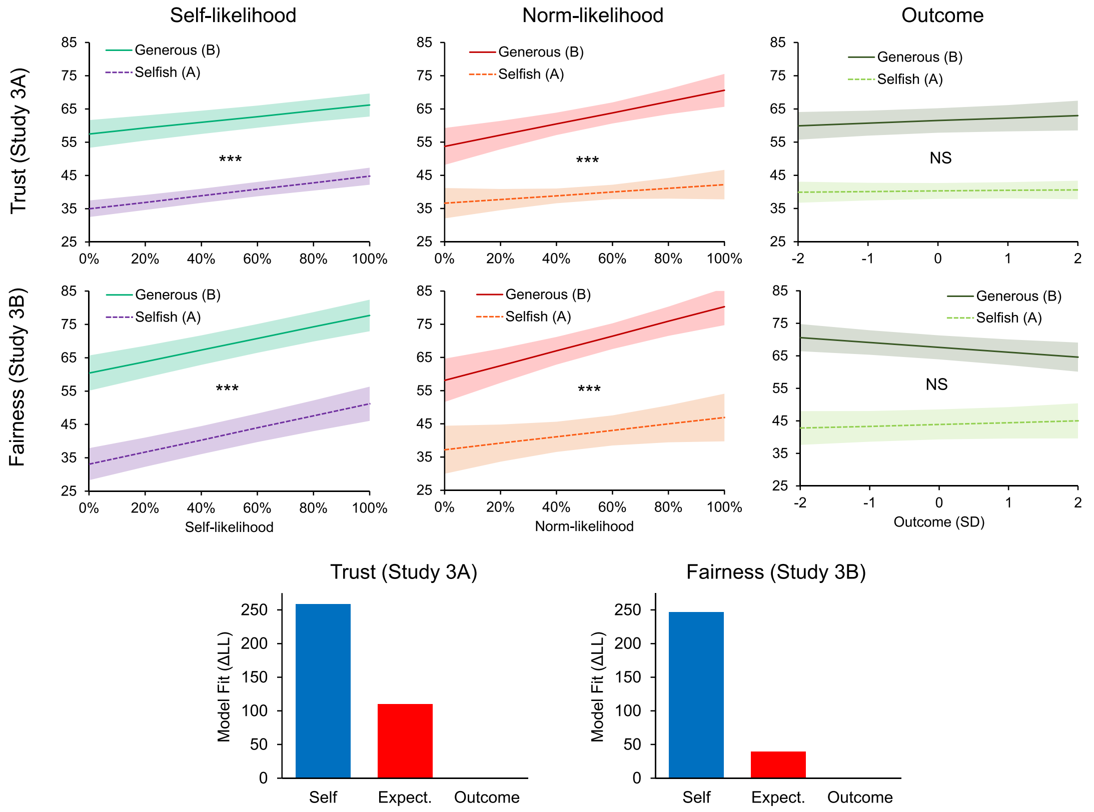

Fig. 5. Self-likelihood best predicted trust and fairness judgments. (Top) The lines represent the effect of *Self*, *Norm*, and *Outcome* on participants' trust judgments (Study 3A); the *Outcome* variable was z-standardized, and the x-axis represents standard deviations (SD) because *Outcome* is not bound to 0–100%. Shaded regions represent 95% confidence intervals. (Middle) The same lines were prepared but with respect to participants' fairness judgments (Study 3B). (Bottom) The bars represent the model fits of regressions predicting each study's dependent using one of these three variables along with the computer's choice (A or B) as predictors. Fit is reported as the log-likelihood difference (ΔLL) relative to a plain model with random intercepts and a fixed effect for the computer's choice (A or B). ***, $p < .001$.

proposed model applies to both character-level and action-level moral judgments. We discuss these results below.

5.1. Possible contexts and Bayesian inference

The present work is the first to probabilistically model moral evaluations as a function of specific possible contexts. However, it is important to note that this idea is rooted in seeds set by older work, such as studies on Attribution Theory (Kelley & Michela, 1980; Ross, 1977). These earlier studies and theoretical papers described how, after observing another person's behavior, individuals assess whether the behavior should be attributed to the person's "dispositions" and/or their "situation." In more recent years, Theory of Mind researchers have formalized these ideas, putting forth probabilistic models to describe how individuals draw inferences about other people's behavior (Baker et al., 2011; FeldmanHall & Shenhav, 2019; Jara-Ettinger, 2019). The present research indeed finds that, after observing another person's behavior, individuals use Bayesian inference to reason about the person's context. In this way, the present research may be seen as bridging work on Theory of Mind to understanding moral and social judgment. However, one notable difference relative to these earlier studies is that they generally also involve modeling how participants draw inferences about other people's traits (e.g., the other person is strong/weak or the other person is generous/selfish). These models presume that inferences about another person's traits occur simultaneously with inferences about the person's context, such that the most likely explanation for someone's behavior is a mix of these elements. Although not tested here, prior beliefs about a person's traits will additionally impact how one reasons about their context, and this point will be discussed further below.

Beyond influencing inferences about possible contexts, it is possible that beliefs about a person's traits themselves also more directly contribute to moral judgment (Carlson, Bigman, Gray, Ferguson, & Crockett, 2022; Uhlmann, Pizarro, & Diermeier, 2015). To an extent, we studied this topic by modeling the Study 3 participants' concerns for inequity and their perceptions of the other players' concerns for inequity (Supplementary Materials 2). These results suggested that participants' trust and fairness judgments were influenced by the absolute difference between participant's equity concerns and their perceptions of others' equity concerns. Although these differences predicted judgments less so than self-likelihood, they speak to how a more complete model of social judgment may incorporate multiple pathways for the assessment of similarity.

Because our goal was to investigate moral judgment during uncertainty, our studies contained salient elements of uncertainty, particularly our economic game designs. However, it is interesting to consider that, in principle, *every* moral judgment carries some uncertainty. This begs the question of whether our proposed model is relevant to *all* forms of social judgment, even when there is no obvious uncertainty present to excuse inappropriate behavior. For instance, suppose an individual is walking and nearly gets hit by an apparent drunk driver. Even if the pedestrian has never driven drunk themselves, will their anger hinge on the likelihood that they too would drive drunk? Although this may sound absurd, it could be true. Perhaps, the pedestrian wonders whether the driver needed to take someone to the hospital for an emergency and nobody else could help. Social judgment would hence depend on how absurd and improbable the circumstances would need to be for the pedestrian to agree that they would behave similarly. Following this logic, the model and mechanisms proposed here could, in principle, be applied to any social judgment. However, more empirical work is needed to test whether this possibility genuinely describes human cognition.

The mechanisms investigated here concerning possible contexts and Bayesian inference may shed light on other questions related to moral judgment. For instance, it is interesting to consider how biases in these inferences may, in turn, produce biases in judgment overall. If someone assumes that most others are selfish, then when viewing uncertain behavior, they may tend to infer possible contexts consistent with selfish behavior, which in turn biases them toward further negative judgments. Conversely, viewing others as mostly good could encourage people to consider possible contexts consistent with this and beget positive judgments.

Considering biases may be particularly relevant for understanding moral judgment concerning an individual's in-group or out-group. Previous studies on social judgment have yielded conflicting findings on the effects of group memberships. Some showed that social violations by in-group members were taken less seriously (Kubota, Li, Bar-David, Banaji, & Phelps, 2013; Mills, Tainsky, Green, & Leopkey, 2018), whereas other studies generated opposing results, finding that violations were taken more seriously when committed by in-group members (Mendoza, Lane, & Amodio, 2014; Wu & Gao, 2018). Although seemingly in direct opposition, these two branches of results can be united by considering possible contexts. When evaluating in-group members, individuals may be biased to ascribe high likelihoods to possible contexts where they would have behaved the same way (Ultimate Attribution Error; Pettigrew, 1979). However, this reasoning is predicated on there being sufficient uncertainty surrounding the in-group members' behavior. If an in-group members' context instead carries little uncertainty, then these types of charitable biases do not have plausible room to manifest. Because they are in-group members, unequivocally inappropriate behavior would be particularly discordant relative to prior impressions. Hence, in these situations where charitable possible contexts are limited, judgment would be harsher. Overall, this reasoning demonstrates the usefulness of constructs developed and investigated here for studying a wider array of topics.

5.2. *Questions and extensions*

One potential criticism of the self-likelihood model and our focus on possible contexts is that this framing of moral judgment may not align with intuition. Everyone can imagine times when they passed judgment on another person while seemingly not giving any consideration to the person's context, let alone a variety of possible contexts. This would be particularly likely for fast judgments—that is, knee-jerk reactions that occur immediately upon observing another person's behavior. This intuition may be seen to challenge our proposed model's scope and suggest that inference about possible contexts is only a minor element of moral judgment.

We have three rebuttals for why this rationale does not dismantle claims about our model's potential importance. (1) Even though individuals may not carefully deliberate possible contexts for every judgment, they may have done so in the past and developed heuristics based on this past reasoning (Gigerenzer, 2008). (2) Consideration of possible contexts may be a stochastic process. When individuals assess the likelihood that they would behave the same way, they may sample from possible contexts with likelihoods that reflect probabilistic inference (e.g., like some decision-making models; Hotaling, 2020). Hence, the self-likelihood mechanisms may differentially impact fast versus slow judgments. (3) Finally, it may simply be the case that inference about possible contexts, comparison to oneself, and integration over this landscape occur below conscious awareness. Altogether, we believe that these different explanations provide a plausible rationale for why the proposed model may be valid, even if it may diverge from intuition.

Finally, we acknowledge that our studies were observational, which generally limits claims about specific mechanisms. Observational research also opens the door to arguments that participants' evaluations and decisions were both driven by some third variable (e.g., an abstract mental representation of "right vs. wrong"). If such is the case, participants may not be truly comparing other people's behavior to their own. However, arguments on the possibility of confounds are limited in explaining our more complex findings supporting the self-likelihood model. Specifically, the Study 2 results comparing $Self_{low}$ versus $Self_{high}$ essentially showed a three-way interaction, whereby participants' overall judgments were influenced by the interplay between (i) the action being evaluated, (ii) their decision-making in a given possible context, and (iii) their views about social norms in said possible context. This type of intricate relationship is difficult to explain in terms of confounds. In addition, even if it is based solely on observational data, observational results deserve explanation and interpretation, and we believe that our self-likelihood model does this succinctly.

6. Conclusion

We investigated moral judgment during uncertainty across four studies, one using vignettes and three using economic games, and we found support for a proposed self-likelihood model of moral judgment. The model posits that, when the context of another person's behavior is uncertain, individuals will consider different possible contexts that could explain the person's behavior. In turn, individuals will consider the overall likelihood that they would behave the same way, and if this likelihood is high, then they will judge the person's behavior more positively. Compared to existing models, the proposed self-likelihood view better predicted participants' judgments in all of our studies. This model may have broad relevance, given that uncertainty is a defining feature of social life and, mechanistically, the proposed model's hypothesized computational pathways parallel those behind other processes involved in social cognition (e.g., Theory of Mind and reciprocity). Hence, we expect that the present research will serve as a robust foundation for further research on these topics and the links between them.

Acknowledgments

This research was carried out in part at the University of Illinois' Beckman Institute for Advanced Science & Technology. During the preparation of this manuscript, P.C.B. was supported by a Predoctoral Fellowship provided by the Beckman Foundation and a Dissertation Completion Fellowship provided by the University of Illinois. F.D. was supported by an Emanuel Donchin Professorial Scholarship in Psychology from the University of Illinois.

Data availability statement

The data and R code needed to reproduce the results for every study in this report are available in a public repository (https://osf.io/h5mea).

Notes

[1] The studies in the present report were not preregistered. However, as will be detailed, all of our results are quite robust (every p-value corresponding to a hypothesized relationship is at least $p < .01$ and generally $p < .001$), and are thus likely replicable.

[2] Inference about possible contexts is the only aspect of Bayesian reasoning examined here (i.e., no analyses were done on how participants form predictions based on prior encounters).

[3] Participants were told that their earnings would depend on their own choices and the decisions of *past* participants. Accordingly, participants were told that their choices would impact the earnings of *future* participants. This design, where participants are influenced by others' past decisions and influence future participants, is common in economic game research (e.g., Bailey, Ruffman, & Rendell, 2013; Chang & Sanfey, 2013).

[4] For instance, suppose a participant is evaluating another player's choice to select Option A. If the participant believes that 20% of people would choose Option A in *Possible Context 1*, and 60% of people would choose Option A in *Possible Context 2*, then the participant can infer that *Possible Context 2* is three times as likely to have been selected than *Possible Context 1*. This is equivalent to simply treating the perceived likelihood of *Possible Context 1* as 0.20 and the likelihood of *Possible Context 2* as 0.60.

References

Bailey, P. E., Ruffman, T., & Rendell, P. G. (2013). Age-related differences in social economic decision making: The Ultimatum Game. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *68*(3), 356–363.

Baker, C. L., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

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.

Bocian, K., Baryla, W., Kulesza, W. M., Schnall, S., & Wojciszke, B. (2018). The mere liking effect: Attitudinal influences on attributions of moral character. *Journal of Experimental Social Psychology*, *79*, 9–20.

Bogdan, P. C., Dolcos, F., Moore, M., Kuznietsov, I., Culpepper, S. A., & Dolcos, S. (2023). Social expectations are primarily rooted in reciprocity: An investigation of fairness, cooperation, and trustworthiness. *Cognitive Science*, *47*(8), e13326.

Bogdan, P. C., Moore, M., Kuznietsov, I., Frank, J. D., Federmeier, K. D., Dolcos, S., & Dolcos, F. (2022). Direct feedback and social conformity promote behavioral change via mechanisms indexed by centroparietal positivity: Electrophysiological evidence from a role-swapping ultimatum game. *Psychophysiology*, *59*(4), e13985.

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.

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

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*(3), 277–284.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359.

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69.

Engelmann, N., & Waldmann, M. R. (2022). How to weigh lives. A computational model of moral judgment in multiple-outcome structures. *Cognition*, *218*, 104910.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*, *123*(3), 434–441.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435.

Fischer-Lokou, J., Martin, A., Guéguen, N., & Lamy, L. (2011). Mimicry and propagation of prosocial behavior in a natural setting. *Psychological Reports*, *108*(2), 599–605.

Gigerenzer, G. (2008). Moral intuition = fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity* (pp. 1–26). Boston Review.

Goldring, M. R., & Heiphetz, L. (2020). Sensitivity to ingroup and outgroup norms in the association between commonality and morality. *Journal of Experimental Social Psychology*, *91*, 104025.

Gollwitzer, A., Martel, C., Bargh, J. A., & Chang, S. W. (2020). Aversion towards simple broken patterns predicts moral judgment. *Personality and Individual Differences*, *160*, 109810.

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.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30–39.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*(5868), 1362–1367.

Hotaling, J. M. (2020). Decision field theory-planning: A cognitive model of planning on the fly in multistage decision making. *Decision*, *7*(1), 20.

Irwin, K., & Horne, C. (2013). A normative explanation of antisocial punishment. *Social Science Research*, *42*(2), 562–570.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.

Jouxtel, J. (2019). Voluntary contributions of time: Time-based incentives in a linear public goods game. *Journal of Economic Psychology*, *75*, 102139.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kawamura, Y., & Kusumi, T. (2020). Altruism does not always lead to a good reputation: A normative explanation. *Journal of Experimental Social Psychology*, *90*, 104021.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology*, *31*(1), 457–501.

Kim, M., Mende-Siedlecki, P., Anzellotti, S., & Young, L. (2021). Theory of mind following the violation of strong and weak prior beliefs. *Cerebral Cortex*, *31*(2), 884–898.

Kim, M., Park, B., & Young, L. (2020). The psychology of motivated versus rational impression updating. *Trends in Cognitive Sciences*, *24*(2), 101–111.

Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the Ultimatum Game. *Psychological Science*, *24*(12), 2498–2504.

Kulesza, W., Chrobot, N., Dolinski, D., Muniak, P., Bińkowska, D., Grzyb, T., & Genschow, O. (2022). Imagining is not observing: The role of simulation processes within the mimicry-liking expressway. *Journal of Nonverbal Behavior*, *46*, 233–246.

Kulesza, W., Dolinski, D., Huisman, A., & Majewski, R. (2014). The echo effect: The power of verbal mimicry to influence prosocial behavior. *Journal of Language and Social Psychology*, *33*(2), 183–201.

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

Lindström, B., Jangard, S., Selbing, I., & Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, *147*(2), 228.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For members only: Ingroup punishment of fairness norm violations in the Ultimatum Game. *Social Psychological and Personality Science*, *5*(6), 662–670.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Mills, B. M., Tainsky, S., Green, B. C., & Leopkey, B. (2018). The Ultimatum Game in the college football rivalry context. *Journal of Sport Management*, *32*(1), 11–23.

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, *95*(1), 76.

Noussair, C. N., & Stoop, J. (2015). Time as a medium of reward in three social preference experiments. *Experimental Economics*, *18*(3), 442–456.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin*, *5*(4), 461–476.

Quiros, J. D. V., Kapcak, O., Hung, H., & Cabrera-Quiros, L. (2021). Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates. *IEEE Transactions on Affective Computing*, *14*(13), 2168–2181.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. *Advances in Experimental Social Psychology*, *10*, 173–220.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70.

Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Vavra, P., Chang, L. J., & Sanfey, A. G. (2018). Expectations in the Ultimatum Game: Distinct effects of mean and variance of expected offers. *Frontiers in Psychology*, *9*, 992. https://doi.org/10.3389/fpsyg.2018.00992

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils you know, than the one you don't: Predictability influences moral judgments of immoral actors. *Journal of Experimental Social Psychology*, *97*, 104220.

Wu, Z., & Gao, X. (2018). Preschoolers' group bias in punishing selfishness in the Ultimatum Game. *Journal of Experimental Child Psychology*, *166*, 280–292.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099–1108.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supplementary Materials