Without his cookies, he's just a monster: A counterfactual simulation model of social explanation

Erik Brockbank* (ebrockbank@stanford.edu), Justin Yang* (justin.yang@stanford.edu), Mishika Govil, Judith E. Fan, Tobias Gerstenberg
Stanford University

Abstract

Everyday reasoning about others involves accounting for why they act the way they do. With many explanations for someone's behavior, how do observers choose the best one? A large body of work in social psychology suggests that people's explanations rely heavily on traits rather than external factors. Recent results have called this into question, arguing that people balance traits, mental states, and situation to make sense of others' actions. *How might they achieve this?* In the current work, we hypothesize that people rely on *counterfactual simulation* to weigh different explanations for others' behavior. We propose a computational model of this process that makes concrete predictions about when people will prefer to explain events based on the actor's *traits* or their *situation*. We test the predictions of this model in an experimental paradigm in which trait and situation each guide behavior to varying degrees. Our model predicts people's causal judgments well overall but is less accurate for trait explanations than situational explanations. In a comparison with simpler causal heuristics, a majority of participants were better predicted by the counterfactual model. These results point the way toward a more comprehensive understanding of how social reasoning is performed within the context of domain-general causal inference.

Keywords: causal reasoning; explanation; attribution; traits.

Commonsense reasoning about the world around us requires accounting for *why* things happened the way they did. Nowhere is this more apparent than in our interactions with other people. How we explain the behavior of others has been a central focus in social psychology for nearly 100 years (Heider & Simmel, 1944), yet incorporating prior findings into AI and applying them to naturalistic scenarios remains a challenge (Lake et al., 2017). One reason for this is the sheer diversity of explanations for others' actions. Consider coming home and finding one's roommate baking a dessert: this could be a thoughtful gesture or merely a chance to use up eggs that will soon go bad. The way we account for people's behavior has consequences for how we treat them and make plans with them (Carlson et al., 2022; Ho et al., 2022); in short, explanations for others' actions are woven into the fabric of our social lives. *Confronted with a range of explanations for others' behavior, how do we choose the best one?*

Inferring traits and mental states from behavior When accounting for others' behavior, observers must decide whether to place the cause *within* the actor or in contextual factors *external* to the actor. Early work exploring this tradeoff found a widespread overreliance on *traits* as the putative cause of others' actions (Gawronski, 2004; Gilbert & Malone, 1995; Ross & Nisbett, 1991). The most common explanation for this finding has been that people simply overlook or ignore the degree to which situation causes behavior (the *fundamental attribution error*; Jones and Harris, 1967; Ross, 1977). People seem to infer traits rapidly and spontaneously (Uleman et al., 2008; Winter & Uleman, 1984), while our ability to analyze the broader context in which somebody acted can be costly and error-prone (Gilbert & Malone, 1995).

However, recent findings have called into question the ubiquity of this trait bias. A meta-analysis of 173 studies documenting the fundamental attribution error found mixed evidence for the effect and suggested that it was moderated by other variables such as social proximity (Malle, 2006). When people read about others performing a wide range of everyday activities, they inferred the actors' intentions and beliefs more readily than traits (Malle & Holbrook, 2012). And cross-cultural studies find a large variance in the prevalence and expression of the bias outside Western cultures (Choi et al., 1999; Knowles et al., 2001; Miyamoto & Kitayama, 2002).

In an effort to reconcile the trait bias with other causes of behavior like situation and mental states, recent work has proposed that the extent to which people infer traits from others' actions is flexible (FeldmanHall & Shenhav, 2019; van Baar et al., 2022). Walker et al. (2022) argue that prior work showing the fundamental attribution error has assumed trait and situation are *deterministic* causes; if traits are instead treated as probabilistically influencing behavior, results observed in classic fundamental attribution error studies arise from a model that makes context-sensitive trait inferences.

Explaining behavior with traits and mental states Prior work suggests that people flexibly infer the traits, mental states, and situational pressures that could have given rise to others' actions. This poses a puzzle: how do we decide which of these factors offers the best *explanation* of their behavior? There is an important conceptual distinction between the *inferences* we draw about others from their actions, and the *explanations* we provide for those same actions (Korman & Malle, 2016). Just as we can infer mental states, traits, or important situational factors on the basis of others' actions, all of these represent plausible *causes* that we might use to explain their behavior; in other words, *explaining others' actions requires selecting from among competing causes*.

How do people solve this problem? In scientific and legal settings, counterfactuals are intimately tied to what makes a

good explanation (Mackie, 1974); meanwhile, formal tools for evaluating counterfactuals have been immensely important for modern AI (Pearl, 2000). However, the degree to which people rely on counterfactuals for everyday causal reasoning is unclear (Galinsky & Moskowitz, 2000; Mandel, 2003; Mandel & Lehman, 1998). Some have argued that counterfactual simulation is essential to any account of human causal reasoning (Kahneman & Miller, 1986; Wells & Gavanski, 1989). A growing body of work exploring causal judgments about *physical* events has found that counterfactual simulations performed with a noisy physics engine accurately predict people's inferences and even their eye movements (Gerstenberg, 2022; Gerstenberg & Stephan, 2021; Gerstenberg et al., 2017). Efforts to extend these findings to social cognition suggest that evaluations of *responsibility* and *blame* involve counterfactual simulation of how an agent would have behaved under different circumstances (Wu et al., 2023). Yet simulating complex alternatives can be costly and difficult, especially when they concern the behavior of others. For this reason, it's been proposed that explanations of others' actions may instead recruit simpler *proxies* for the relevant counterfactuals (Lipe, 1991).

Our approach In the current work, we hypothesize that to explain others' actions, people simulate whether the outcome of those actions would have differed if the causes had been otherwise. We propose a computational model that compares *trait* and *situation*-based explanations of behavior through counterfactual simulation of these variables (Figure 1 Top). We test the predictions of the model in an experimental paradigm that allows for fine-grained control over the role of trait and situation in downstream outcomes. While prior work has argued that people eschew counterfactual simulation for simpler approximations, our model predicts individual judgments about the cause of others' actions better than alternatives that rely on such approximations. These results provide suggestive evidence for conditions in which social explanations draw on domain-general causal reasoning.

Study environment: manipulating trait and situation as causes of behavior

We examine social reasoning in an environment where agents' traits and how much behavior is constrained by the situation are quantifiable and can be flexibly manipulated. Participants were shown a series of 10x10 grid worlds whose cells were populated with 10 berry trees at randomly selected locations (Figure 2). Each berry tree contained between one and nine berries (rewards were sampled uniformly). In every grid, a number of *mystery trees* had unknown rewards (each tree's visibility was a binomial variable with probability $p = .75$ of being visible). One of two *farmer agents* harvested as many berries as possible in the grid. Agents could only move up to 10 steps, so they chose their paths carefully.

In this environment, the farmers' planning is formalized as a *Markov Decision Process* (MDP) (Bellman, 1957) in which the best path is the one that maximizes expected reward within the steps available. The two farmers differed in how they computed expected reward. The *optimist* farmer, HOPE, expected mystery trees to have a high reward of eight berries. Meanwhile, PRUDENCE, the *pessimist* farmer, expected mystery trees to have a symmetrically low reward of only two berries.[1] As a result, the agents planned their routes in ways that exhibited systematic, trait-like behavior.

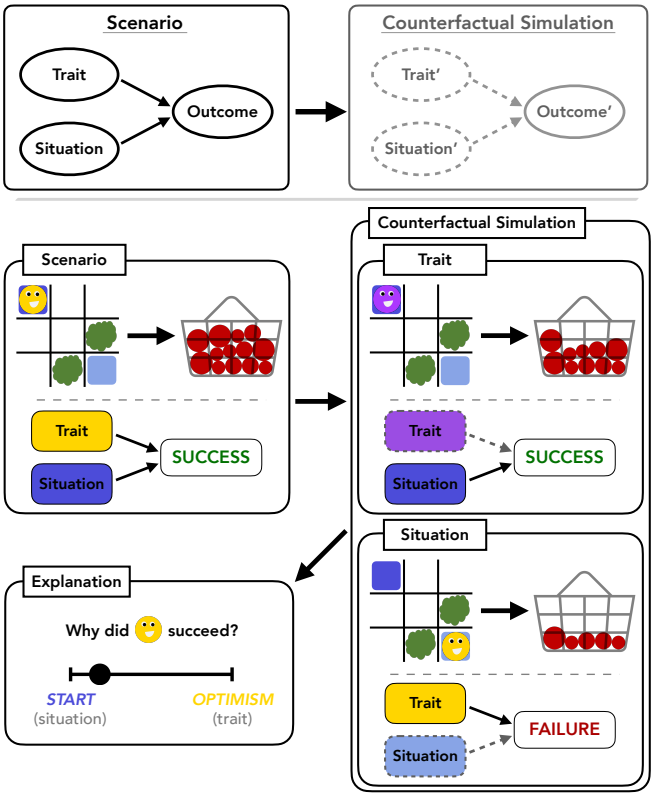

Figure 1: ***Model Schematic***. (Top) We hypothesize that people choose the best explanation for others' actions by simulating how things would have turned out if *trait* or *situational* causes differed. (Bottom) Our experiment probes judgments about the outcomes of two farmers that vary in their underlying *optimism* (trait) and harvest from two *start locations* (situation). Our model predicts these judgments via counterfactual simulations of the farmer's trait and situation.

Critically, the farmers' traits were not the only factors contributing to the outcome of their harvests. To make the causal role of the *situation* concrete, the farmers always began harvesting from one of two squares on opposite corners of the grid. As a result, their start location sometimes positioned them close to high reward trees and sometimes placed them at a disadvantage. Across a range of grid world environments, participants were prompted to evaluate the degree to which the agents' harvest outcomes were caused by their *traits* (optimism or pessimism) and their *situation* (start location).

[1]These expectations can be formalized as *beta binomial* distributions over the range of possible rewards, but for the current experiment, all that matters is their means of two and eight.

Counterfactual simulation model of trait and situational causes

We propose a model of social explanation in which observers reason about the causes of a farmer's harvest outcome through counterfactual simulation (Figure 1). In each grid, the farmer's harvest was considered a *success* if they collected 20 or more berries within the 10 steps allotted.[2] To estimate the causal role that the farmer's trait played in their outcome, the model simulates how often the outcome would have differed if the agent harvesting had the *other agent's trait*. For example, in a plot where the *optimistic* farmer *succeeded*, the model simulates how often they would have failed if they instead harvested like the pessimistic farmer. To estimate the role that the situation played in a harvest outcome, the model simulates how the outcome might have differed if the farmer had instead started from the *other start location*. The farmer's start location is just one situational cause among many (e.g., how were rewards distributed across nearby trees?), but allows us to simulate reasoning about situational causes without introducing more complex inferences.

Counterfactual simulation To estimate the counterfactual probability of the farmer's outcome changing from the other start location or with the other farmer's trait level, the model samples k paths in proportion to the probability of these paths in the relevant counterfactual (e.g., the probability of different paths taken by the agent from the other start location). The probability of these counterfactual paths is computed using the softmax of the *expected* reward on each possible path; thus, the paths each agent would have taken reflect their underlying optimism or pessimism. For each sampled path, the model combines the rewards from visible trees on the path with mystery tree reward values sampled from the true uniform distribution. The counterfactual probability of the agent's outcome changing is the proportion of the k sampled paths in which this reward estimate *would have changed* the original outcome. We use $k = 1{,}000$ sampled paths in all reported results. These paths are sampled in proportion to a log normal distribution centered at 10 steps and truncated at eight and 12 with variance σ_{steps} to accommodate the possibility that counterfactual simulation might sometimes "mis-count" the steps an agent takes. In addition, the model samples visible tree rewards in each counterfactual path from a log normal distribution centered at the tree's true value with variance σ_{reward} to allow for the possibility that evaluation of how an agent *would have fared* may rely on noisy counts.

Our model predicts that judgments about whether an agent's trait or start location *caused* their outcome are a function of these simulated counterfactuals (i.e., the estimated probability that their outcome would have changed if their trait or start location had been otherwise). Our model further predicts that people choose the *best* explanation for an outcome based on the relative strength of each cause.

[2]This cutoff was a salient number close to the median total harvest in a sample of 1,000 simulated grid worlds; see methods.

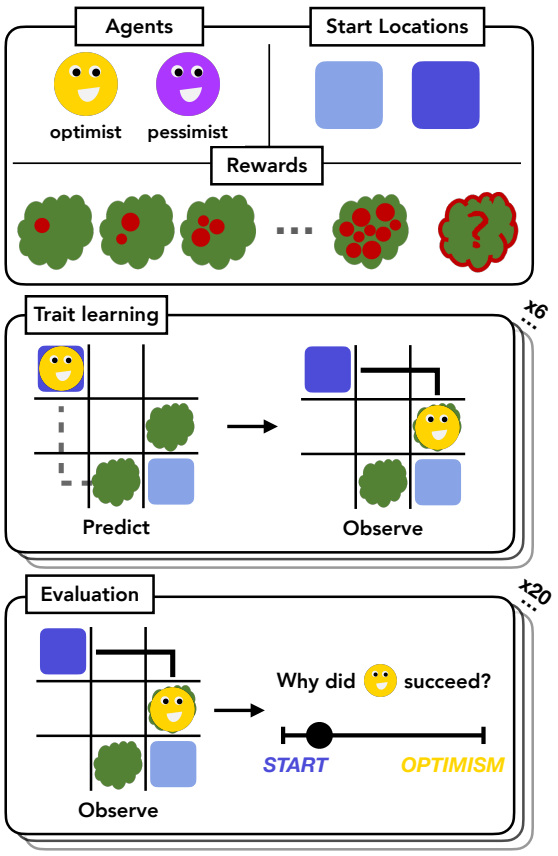

Figure 2: ***Experiment Overview***. Two farmers, one *optimistic* and the other *pessimistic*, try to harvest as many berries as possible from different grids on their farm. Participants completed six *learning trials* to understand how the farmers chose their paths, followed by 20 *evaluation trials* in which they watched a farmer harvest a new plot and answered one of five *counterfactual* or *causal* questions about the farmer's harvest.

Experiment

The methods and analyses in the current study, along with a more comprehensive description of experiment stimuli, were preregistered on the OSF at https://osf.io/d48jk.[3]

Participants

Participants were recruited on Prolific. We recruited 30 participants in each of five between-subjects conditions for a total of $N = 150$ participants (*age*: median = 34, range = 20-74; *gender*: 88 female, 58 male, 4 non-binary; *race*: 101 white, 24 black/African American, 14 Asian, 10 multiracial). All participants were native English speakers residing in the US. Participants were paid $5 for an estimated 25 minutes to complete the study (*mean completion time*: 18.0 mins). This study followed the Stanford University IRB protocol and all participants provided informed consent.

[3]Experiment code and analyses are available on github at https://github.com/cicl-stanford/action_abstraction_cogsci2024.

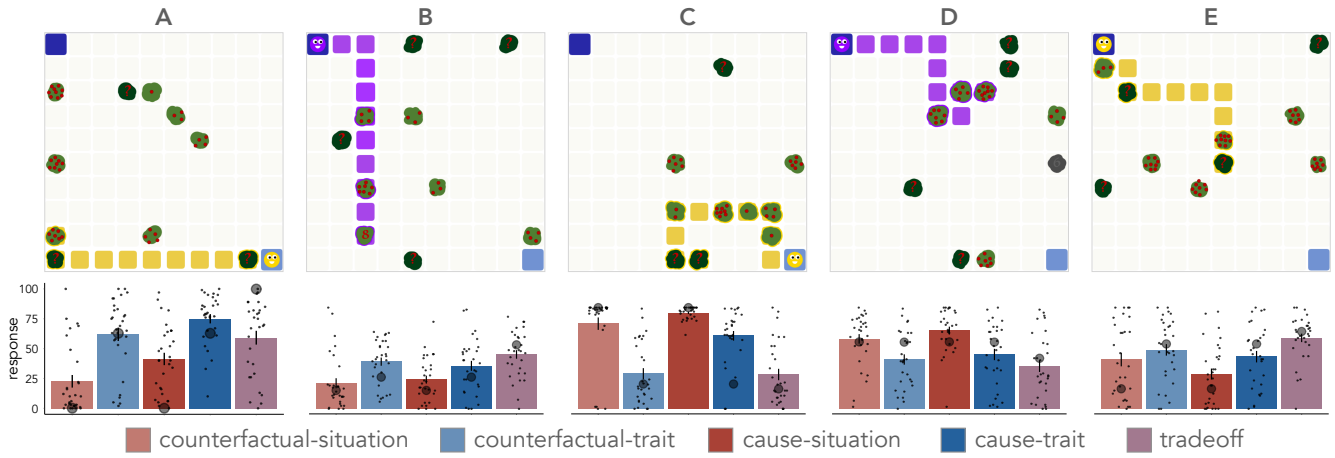

Figure 3: ***Sample trials***. Bar plots show average responses with individual responses overlaid. Large dots show model predictions. All error bars are standard error of the mean (SEM). (A)-(B): Participants and the model both predicted that the farmer's outcomes were caused by their respective traits. (C)-(D): People and the model attributed the farmers' outcomes to their start location. (E): Participants and the model exhibited a qualitatively similar pattern of uncertainty about the causes.

Stimuli

Even with existing constraints on the size of the grid, the number of trees, their possible reward values, and their visibility, the number of possible grid world trials is practically infinite (roughly 14 trillion based on possible tree locations alone). To select individual trials for this experiment, we procedurally generated a sample distribution of 1,000 grid worlds with an assigned agent. For the 1,000 sample $(gridworld, agent)$ environments, we simulated the agent's most likely path, along with relevant counterfactuals such as their outcome from the other start location and with the other agent's trait level. We then selected 20 to serve as experiment trials by applying a set of predetermined criteria (for example, a balanced distribution of assigned agent, harvest outcome, and counterfactual outcomes; see OSF preregistration).

Procedure

Participants were first shown instructions detailing the study environment and the farmers. They were told that some trees in each plot would contain visible rewards and that others would be unknown to them and the farmer harvesting the plot. They were told that the two farmers had different expectations about the *mystery trees*; HOPE was *optimistic* and routinely expected mystery trees to contain many berries, while PRUDENCE was *pessimistic* and expected mystery trees to have very few berries. The experiment consisted of a *learning phase* and an *evaluation phase* (Figure 2).

In the *learning phase*, participants completed six trials in which their goal was to familiarize themselves with the way each farmer harvested. Participants were first shown the grid world and the farmer's starting location and instructed to *click the trees in the order they thought the farmer would harvest them*. As they clicked, the farmer's path from one tree to the next was animated and a counter indicated how many of their 10 steps remained. When there were no more trees accessible from a predicted tree, participants could submit their prediction. Participants could *undo* their predictions by clicking previous trees and revise until they were satisfied with their answer. Upon submitting their prediction, participants were shown the farmer's true path animated over the predicted path and given feedback on the trees they correctly predicted.

After the learning trials, participants completed a comprehension check consisting of two questions in which they were shown a novel grid world (one for each farmer) and three possible paths. Participants could not proceed to the *evaluation phase* until they had selected the correct path in both questions. Participants then completed 20 *evaluation trials*. In these trials, participants first watched an animation of the farmer's path through a novel grid world; they were then shown the outcome of this harvest and prompted to evaluate the outcome with a slider question.

Across five between-subjects conditions, participants were shown different prompts that tested predictions of the counterfactual simulation model. Those in the ***counterfactual situation*** condition were asked to indicate how strongly they agreed that the farmer's outcome would have differed if they had started on the other start location, using a slider with "Not at all" at one end and "Strongly" at the other. Participants in the ***counterfactual trait*** condition were instead asked to indicate on the same slider how strongly they agreed that the farmer's outcome would have differed if they were as optimistic or pessimistic as the other farmer. Participants in the ***causal situation*** condition indicated how strongly they agreed that the farmer's outcome was caused by their start location. Participants in the ***causal trait*** condition indicated how strongly they agreed that the farmer's outcome was caused by their optimism or pessimism. Finally, participants in the ***causal tradeoff*** condition were shown a slider

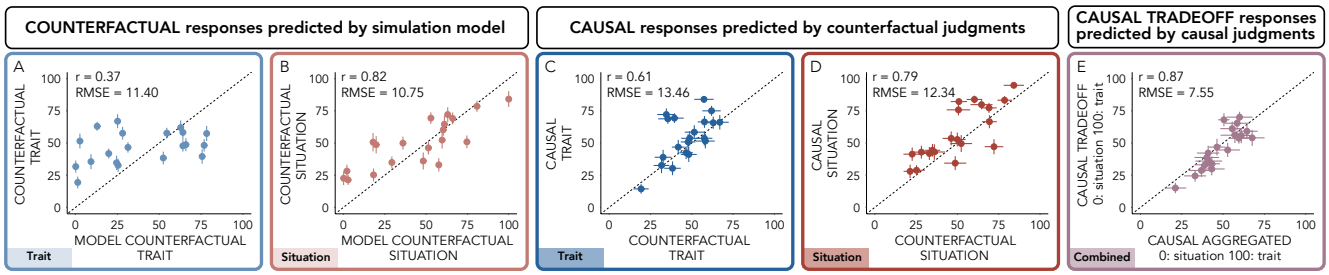

Figure 4: ***Experiment results.*** (A)-(B): Relationship between participants' counterfactual judgments and model predictions. (C)-(D): Relationship between participants' counterfactual judgments and causal evaluations. (E): Relationship between causal tradeoff responses and causal evaluations normalized to predict this tradeoff. Error bars are standard error of the mean (SEM).

with "Start location" at one end and the farmer's trait at the other and asked *why* the farmer succeeded or failed (Figure 2).

The counterfactual simulation model makes concrete predictions for each of these conditions. We compare participants' responses to these predictions to evaluate whether their behavior is consistent with using counterfactual simulation to support causal reasoning about the farmers' behavior.

Results

Participants acquired an accurate predictive model of each farmer before the evaluation trials. Prediction accuracy on the learning trials was high overall ($mean = 75.8\%$, $sd = 33.2\%$) and remained stable over the course of these trials; a mixed effects model fit to prediction accuracy with random intercepts for both subjects and trial stimuli found no effect of trial index ($\chi^2(1) = 0.40$, $p = .70$). There was also no difference in accuracy across conditions ($\chi^2(4) = 0.98$, $p = .45$). The comprehension check after the learning trials was meant to provide an additional indication of whether participants understood each farmer's harvesting behavior. 74.7% of participants answered both comprehension check questions correctly on their first attempt and 92.0% completed them with two retries or fewer, leaving only 12 participants who required three or more retries. The results below focus on behavior in the subsequent *evaluation trials* (see Figure 3).

Counterfactual judgments are consistent with simulations of situation more than trait

We first assess whether participants' counterfactual judgments can be predicted by our counterfactual simulation model (Figure 4A-B). The model has three free parameters: the softmax temperature τ for sampling counterfactual paths, and the log normal variance for the step count (σ_{steps}) and visible berry count (σ_{reward}) distributions. We fit these parameters using a grid search that minimized the squared error of average counterfactual judgments on each trial. All results reported here use the best-fitting parameter values. The simulation model's estimated probability of the *start location* changing trial outcomes was strongly correlated with judgments in the ***counterfactual situation*** condition ($r = 0.82$, $p < .001$). Model predictions for the probability of the farmer's *trait* changing their outcome were positively correlated with responses in the ***counterfactual trait*** condition but were not significant ($r = 0.37$, $p = .11$).

To better characterize the simulation model's prediction accuracy across counterfactual conditions, we calculated the reliability of participants' own counterfactual judgments in each condition. Participants were split into equal halves and the average judgments on each trial were correlated across split halves then corrected using the Spearman-Brown prediction formula (Rouder et al., 2019) to estimate the upper bound on our model's correlation with human responses. We repeated this process for 1,000 random split halves in each condition. Both counterfactual conditions had similar reliability levels (***counterfactual situation*** mean split-half reliability: 0.93, $SD = 0.03$; ***counterfactual trait*** mean split-half reliability: 0.87, $SD = 0.05$). The high split half correlations in both conditions suggests that participants made reliable counterfactual judgments about trait and situation across trials; a substantial amount of the variance in *trait* counterfactuals therefore remains unexplained by our simulation model.

Causal reasoning reflects the degree to which agent traits and situation made a difference

The counterfactual simulation model assumes that causal judgments rely on counterfactual simulation of the possible causes. We therefore test the degree to which participants' causal judgments in the ***causal trait*** and ***causal situation*** conditions can be directly predicted by their counterfactual estimates. Since our model does not strongly capture counterfactual judgments across conditions, we expect participants' own judgments in the ***counterfactual trait*** and ***counterfactual situation*** conditions to better explain their causal judgments. Responses in the ***causal trait*** and ***causal situation*** conditions were significantly correlated with responses in the corresponding counterfactual conditions (*situation*: $r = 0.79$, $p < .0001$; *trait*: $r = 0.61$, $p = .004$; see Figure 4C-D). The relationship between the *simulation model's* counterfactual estimates and participants' causal responses was weaker, though the model was significantly correlated with causal situation judgments ($r = 0.73$, $p < .001$) and positively correlated with causal trait judgments ($r = 0.32$, $p = .17$).

Causal explanations reflect the strength of the underlying causes

Our counterfactual simulation model predicts that to choose among competing explanations for the farmer's outcome, people weigh the strength of each of the potential causes. Participants' normalized responses from the two causal conditions were closely aligned with responses on matched trials in the ***causal tradeoff*** condition ($r = 0.87$, $p < .0001$; Figure 4E). This suggests that answering the *why* question in this condition involves comparing the strength of each cause.

Given the relationship between participants' counterfactual judgments and their causal judgments, we test whether causal tradeoff responses can be directly predicted by participants' *counterfactual* inferences. There was a significant correlation between participants' normalized counterfactual judgments and causal tradeoff responses ($r = 0.61$, $p = .004$). Finally, the *simulation model's* counterfactual estimates, when similarly normalized, were significantly correlated with causal tradeoff responses but to a lesser degree ($r = 0.51$, $p = .02$).

Model comparison

In the previous results, we found evidence that causal judgments about the farmers' outcomes relied on counterfactual reasoning. Here, we consider an alternative, that participants' causal judgments were based on simpler *approximations*. We fit a linear mixed effects regression to responses in the two causal conditions with random intercepts for each participant. Our **counterfactual model** used mean responses for each trial in the corresponding counterfactual conditions as predictors. Meanwhile, our **heuristic model** used simple causal heuristics in place of counterfactual simulation. First, we used the interaction between outcome and agent to predict causal *trait* inferences; this is consistent with mere covariation between outcome and trait providing evidence for the causal strength of each trait with respect to a given outcome (Kelley, 1973; Lipe, 1991). To approximate the causal role of *start location*, we used the interaction between outcome and the average expected reward of trees in the trial, weighted by their Manhattan distance from the farmer (using a discount factor $\gamma = 0.9$ for each step). This visual ensemble (Bauer, 2015) will be larger when there are high reward trees close to where the farmer started and smaller when there are few high rewards nearby, thereby offering a simple approximation for whether the farmer's start location was beneficial.

Models were fit using `brms` (Bürkner, 2017) and compared according to their *estimated log predictive density* in cross-validation (Vehtari et al., 2017). Results are shown in Table 1. We calculate the overall fit across participants, as well as the number of individual participants best fit by each model. Overall, causal *trait* judgments were slightly better predicted by the heuristic model than by counterfactual judgments, while causal *situation* judgments were better accounted for by the counterfactual model.[4] For both trait and situation, the majority of individual participants were better described by the counterfactual models.

Table 1: Overall fit (Δ elpd: lower values are a worse fit) and individual subjects best fit by each model.

Model	Δ elpd (se)	n best
Trait models		
Heuristic	0 (0)	10
Counterfactual	-9.3 (13.7)	20
Situation models		
Counterfactual	0 (0)	23
Heuristic	-24.2 (15.2)	7

Discussion

In the current work, we investigate how people choose the best explanation for others' behavior. We hypothesize that people rely on *counterfactual simulation* to weigh the possible causes of the outcome they've observed. To test this hypothesis, we propose a computational model of causal reasoning based on counterfactual simulation of *trait* and *situational* causes. We evaluate predictions of the model in a rich experimental paradigm in which both trait and situation contribute to observed behavior to varying degrees.

Results provide suggestive evidence that people relied on counterfactual simulation to explain observed outcomes. First, our simulation model closely approximates participants' counterfactual judgments about *situational* variables, but does not correlate significantly with their *trait* counterfactuals. Second, when predicting participants' causal judgments with matched counterfactuals alongside simpler trait and situation *heuristics*, the counterfactual models provide a better account of participants' responses at the individual level, but model comparison for aggregate responses produced mixed results. In short, we find stronger evidence for counterfactual simulation when evaluating situational causes relative to traits of the actors themselves.

Beyond resolving these inconsistencies, more work is needed to address the broader question of how people explain others' behavior in everyday settings. First, an important direction for future work will be testing our account in more naturalistic conditions. The current task employs a simplified version of trait and situational causes, and presents this information to participants directly at the outset. By making the agents' mystery tree reward estimates more graded and responsive to their environment, their behavior may reflect a more intuitive notion of optimism or pessimism. Future work should also consider other traits closer to those observed in recent empirical work on trait evaluations (Lin & Thornton, 2023) and present participants with the challenge of inferring trait and situational causes on their own, rather than being

[4] Results published in *Proceedings of the 46th Annual Meeting of the Cognitive Science Society* incorrectly reported in Table 1 that the heuristic model was a better overall fit to both trait and situation judgments.

provided with them. Finally, a critical avenue for extending the current findings lies in considering causes of behavior beyond trait and situation, such as the tendency to explain others' behavior using mental states like beliefs, desires, and goals (Jara-Ettinger et al., 2020; Korman & Malle, 2016).

Acknowledgments

EB and JEF were supported by an ONR Science of Autonomy Award. JEF received additional support from NSF CAREER #2047191. TG was supported by grants from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and the Cooperative AI Foundation.

References

Bauer, B. (2015). A selective summary of visual averaging research and issues up to 2000. *Journal of Vision*, *15*(4), 14–14.

Bellman, R. (1957). A markovian decision process. *Journal of mathematics and mechanics*, 679–684.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

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

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological bulletin*, *125*(1), 47.

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

Galinsky, A. D., & Moskowitz, G. B. (2000). Counterfactuals as behavioral primes: Priming the simulation heuristic and consideration of alternatives. *Journal of Experimental Social Psychology*, *36*(4), 384–409.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European review of social psychology*, *15*(1), 183–217.

Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B*, *377*(1866), 20210339.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological science*, *28*(12), 1731–1744.

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*, *216*, 104842.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological bulletin*, *117*(1), 21.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959–971.

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.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of experimental social psychology*, *3*(1), 1–24.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological review*, *93*(2), 136.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

Knowles, E. D., Morris, M. W., Chiu, C.-y., & Hong, Y.-y. (2001). Culture and the process of person perception: Evidence for automaticity among east asians in correcting for situational influences on behavior. *Personality and social psychology bulletin*, *27*(10), 1344–1356.

Korman, J., & Malle, B. F. (2016). Grasping for traits or reasons? how people grapple with puzzling social behaviors. *Personality and Social Psychology Bulletin*, *42*(11), 1451–1465.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*, e253.

Lin, C., & Thornton, M. (2023). Evidence for bidirectional causation between trait and mental state inferences. *Journal of Experimental Social Psychology*, *108*, 104495.

Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, *109*(3), 456.

Mackie, J. L. (1974). *The cement ofthe universe*. Oxford: Clarendon Press.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological bulletin*, *132*(6), 895.

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? the likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102*(4), 661.

Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, *132*(3), 419.

Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General*, *127*(3), 269.

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of socially constrained behavior. *Journal of personality and social psychology*, *83*(5), 1239.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (pp. 173–220, Vol. 10). Elsevier.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. McGraw-Hill.

Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). Why most studies of individual differences with inhibition tasks are bound to fail.

Uleman, J. S., Adil Saribay, S., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, *59*, 329–360.

van Baar, J. M., Nassar, M. R., Deng, W., & FeldmanHall, O. (2022). Latent motives guide structure learning during adaptive social choice. *Nature Human Behaviour*, *6*(3), 404–414.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Walker, D., Smith, K. A., & Vul, E. (2022). Reconsidering the "bias" in "the correspondence bias". *Decision*, *9*(3), 263.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of personality and social psychology*, *56*(2), 161.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? evidence for the spontaneousness of trait inferences. *Journal of personality and social psychology*, *47*(2), 237.

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2023). A computational model of responsibility judgments from counterfactual simulations and intention inferences. In M. B. Goldwater, F. Anggoro, B. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.