A computational model of responsibility judgments from counterfactual simulations and intention inferences

Sarah A. Wu[1], Shruti Sridhar[2], Tobias Gerstenberg[1]
{sarahawu, shrutisr, gerstenberg}@stanford.edu
[1]Department of Psychology, Stanford University, USA
[2]Department of Computer Science, Stanford University, USA

Abstract

How responsible someone is for an outcome depends on what causal role their actions played, and what those actions reveal about their mental states, such as their intentions. In this paper, we develop a computational account of responsibility attribution that integrates these two cognitive processes: causal attribution and mental state inference. Our model makes use of a shared generative planning algorithm assumed to approximate people's intuitive theory of mind about others' behavior. We test our model on a variety of animated social scenarios in two experiments. Experiment 1 features simple cases of helping and hindering. Experiment 2 features more complex interactions that require recursive reasoning, including cases where one agent affects another by merely signaling their intentions without physically acting on the world. Across both experiments, our model accurately captures participants' counterfactual simulations and intention inferences, and establishes that these two factors together explain responsibility judgments.

Keywords: responsibility; counterfactual simulation; theory of mind; causal attribution; social inference.

Introduction

Responsibility attributions are ubiquitous and consequential in our everyday lives. From watching dashcam or bodycam footage, people can infer who is at fault in a social situation and why. What underlies people's remarkable ability to make rapid, intuitive responsibility judgments about others' social interactions? Prior work has identified two cognitive processes that drive responsibility judgments: a *causal attribution* about the role a person played in bringing about the outcome, and a *mental state inference* that is informed by the person's actions (Gerstenberg et al., 2018; Langenhoff et al., 2021; Sosa et al., 2021; Carlson et al., 2022). However, these studies do not provide concrete accounts of the mechanisms through which these two processes connect to responsibility judgments, particularly for social interactions that unfold over time. Here, we bridge work on causal attribution and mental state inference to provide a unified computational model of responsibility judgments.

Causal attribution

One way of capturing a person's causal role in a situation is by considering what would have happened in a counterfactual scenario in which they hadn't been there, or had acted differently (Lewis, 1973; Pearl, 2000; Chockler & Halpern, 2004; Halpern & Pearl, 2005; Lagnado et al., 2013; Wu & Gerstenberg, 2023). However, little work has investigated the actual cognitive process by which people simulate counterfactuals involving agents. The process underlying simulation in the physical domain, in contrast, has been elucidated in more detail. The Counterfactual Simulation Model (CSM) developed by Gerstenberg et al. (2021) generates counterfactual scenarios using a physics engine that approximates people's intuitive understanding of physical principles (Gerstenberg & Tenenbaum, 2017). The CSM predicts that people judge an object's causal role by comparing what happened with what would have happened if the object hadn't been there.

Sosa et al. (2021) applied the CSM to moral scenarios by modeling agents who exhibited simple behaviors programmed using a physics engine. But agents in reality are governed by their mental states that, together with the principle of rational action (Dennett, 1987), dictate how they should act in order to achieve their goals most efficiently (Jara-Ettinger et al., 2016; Netanyahu et al., 2021). In prior work, we integrated agentive planning into the CSM to model causal judgments about a single agent in pursuit of a physical goal (Wu et al., 2022). Here, we extend our previous work to multiple agents, and model outcomes that result from social goals such as helping or hindering (see Ullman et al., 2009).

Mental state inference

The mental states of others are usually hidden. However, people can infer them from observable actions using their intuitive theory of mind, a process that has been approximated as Bayesian inverse planning or inverse reinforcement learning in Markov Decision Processes (MDPs) and related formalisms (e.g. Baker et al., 2009, 2017; Ullman et al., 2009; Shu et al., 2020; Zhi-Xuan et al., 2020; Jara-Ettinger, 2019). Here, we focus on inferring an agent's *intention* (Kleiman-Weiner et al., 2015), which is critical for assigning them responsibility (e.g. Lagnado & Channon, 2008).

Computational framework

In this paper, we develop a computational model of responsibility judgments that combines mechanisms of counterfactual simulation and intention inference. Importantly, we apply the same generative model of agent behavior to both simulate counterfactual scenarios involving agents and infer those agents' mental states. While counterfactual simulation is sufficient to explain causal judgments in the physical domain, here we show that intentions additionally matter for respon-

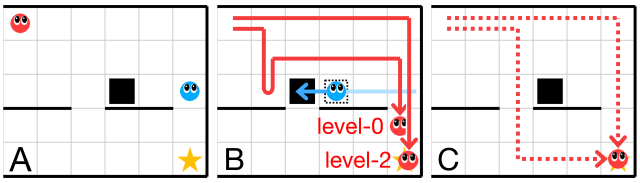

Figure 1: **Example**. (A) RED and BLUE in the environment at timestep $t = 0$. (B) BLUE pushes the box to the left. A level-0 RED backtracks and fails, while a level-2 RED correctly infers BLUE's intent to hinder and succeeds. (C) Some of RED's possible counterfactual paths if BLUE hadn't been there.

sibility judgments, and we demonstrate an implementation of counterfactual simulation for responsibility. We now describe our agents, environment, and models in turn.

Agents

Our setting is a grid world in which agents and objects can interact (Figure 1A). On each timestep, agents can move up, down, left, right, or stay in place, but cannot move through walls or boxes. One agent, RED, has a physical goal of reaching a star in 10 timesteps. If they run out of time, then they fail. Another agent, BLUE, has a social goal of helping or hindering RED. BLUE has the ability to push or pull boxes around. Using an approach similar to level-k reasoning or cognitive hierarchy (Wright, 2010), we model two types of RED: a level-0 RED who plans only towards their physical goal, and a level-2 RED who additionally infers and plans around BLUE's intentions (Figure 1B). This agent can, for example, try to avoid BLUE or wait for them to take some action. BLUE is level-1 and always assumes a level-0 RED.

Environments

Formally, our setting can be represented as a set of Social MDPs (see Tejwani et al., 2021, for the general formulation). The Social MDP M^l_i for agent $i \in \{\mathrm{R}, \mathrm{B}\}$ at level l is the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \chi_i, g_i, \mathcal{R}^l_i, \gamma \rangle$, where $\mathcal{S}$ is the state space of all states $s \in \mathcal{S}$; $\mathcal{A}$ is the joint action space of all pairs of actions a_R, a_B; $\mathcal{T} = \mathcal{T}(s' \mid s, a_\mathrm{R}, a_\mathrm{B})$ dictates the transition probabilities; χ_i is agent i's social goal, g_i is their physical goal; $\mathcal{R}^l_i$ is their reward function, and $\gamma \in (0,1)$ is a reward discount factor.

The level-0 RED has no social goal, so $\chi_\mathrm{R} = 0$. Their reward function is static and depends on their physical goal of reaching the star, along with their action cost $c(a_R)$:

$$\mathcal{R}^0_\mathrm{R}(s, a_\mathrm{R}, a_\mathrm{B}, g_\mathrm{R}) = r_\mathrm{R}(s, a_\mathrm{R}, g_\mathrm{R}) - c(a_R).$$

The level-1 BLUE has no physical goal, so $g_\mathrm{B} = 0$. Their social goal is instantiated as a scaling factor that transforms their estimate of level-0 RED's reward into their own reward:

$$\mathcal{R}^1_\mathrm{B}(s, a_\mathrm{R}, a_\mathrm{B}, g_\mathrm{R}) = \chi_\mathrm{B} \cdot \widetilde{\mathcal{R}}^0_\mathrm{R} - c(a_B) + f(\chi_\mathrm{B}).$$

Here, we set $\chi_\mathrm{B} = 0.5$ for a helping BLUE or $\chi_\mathrm{B} = -0.5$ for a hindering BLUE. To add gradation to BLUE's intentions, we define a supplemental reward term $f(\chi_\mathrm{B})$, which is a function of the change in the number of shortest paths available to RED and the change in the length of the shortest path. Increasing path availability and decreasing path length both reward a helping BLUE, but penalize a hindering BLUE.

BLUE's estimate of a level-0 RED's reward function is $\widetilde{\mathcal{R}}^0_\mathrm{R} = r_\mathrm{R}(s, a_\mathrm{R}, g_\mathrm{R}) - c(a_\mathrm{R})$. Here, RED's physical goal g_R is assumed to be known, and does not need to be estimated.

Finally, since the level-2 RED only has a physical goal, its reward function is identical to that of a level-0 RED agent:

$$\mathcal{R}^2_\mathrm{R}(s, a_\mathrm{R}, a_\mathrm{B}, g_\mathrm{R}) = r_\mathrm{R}(s, a_\mathrm{R}, g_\mathrm{R}) - c(a_\mathrm{R}).$$

We now describe how to plan for these settings and, in particular, how to infer BLUE's social goal (i.e. their intention).

Generative model

Our generative model solves the pairwise Social MDPs M^0_R, M^1_B, and M^2_R. Each trial features a RED who is either level-0 or level-2, and a BLUE who is level-1. To solve M^0_R and M^1_B, the model uses a Q function for each agent i at level l:

$$Q^l_i(s, a_i, g_i, \chi_i) = \sum_{s'} \mathcal{T}(\cdot)\left[\mathcal{R}^l_i(\cdot) + \gamma V^l_i(s', g_i, \chi_i)\right],$$

where V^l_i is the respective value function.

Agents iteratively compute their Q functions and deterministic policies (see Tejwani et al., 2021, for the full algorithm and general formulation of value functions). The level-0 RED plans as in a simple MDP, and the level-1 BLUE plans assuming the level-0 RED's physical goal is known.

To solve M^2_R, the level-2 RED continuously maintains estimates of level-1 BLUE's social goal χ_B. At timestep $t = 0$, their belief $p(\widetilde{\chi}^0_\mathrm{B})$ is initialized to the uniform distribution. On each timestep t thereafter, they perform a Bayesian update:

$$p\left(\widetilde{\chi}^t_\mathrm{B} \mid s^{t-1}, a^{t-1}_\mathrm{B}\right) \propto p\left(a^{t-1}_\mathrm{B} \mid s^{t-1}, \chi_\mathrm{B}\right) p(\widetilde{\chi}^{t-1}_\mathrm{B}).$$

They predict the level-1 BLUE's social policy using their Q function, which they must solve, under a softmax:

$$p\left(a_\mathrm{B} \mid s^t, \widetilde{\chi}_\mathrm{B}\right) \propto \exp\left(\beta \cdot Q^1_\mathrm{B}(s, a_\mathrm{B}, \widetilde{\chi}_\mathrm{B})\right).$$

The softmax accounts for occasional non-optimal actions via the parameter β, which captures an agent's level of randomness while acting. The level-2 RED plans by using Monte Carlo methods to sample possible actions a_B that BLUE could take, and then selecting an optimal path in light of them.

In all experiments, we used an action cost of $c = 1$ for moving in the grid, and $c = 2$ for pushing or pulling a box (for BLUE only). Solving these Social MDPs generates policies for RED and BLUE within the bounds of the grid world and their respective levels of reasoning. These policies approximate people's intuitive theories of how agents interact based on their mental states, capacities, and situational constraints.

Responsibility model

To predict how responsible each agent is for the outcome in each scenario, our responsibility model uses the generative model to infer intentions and to implement counterfactual simulations. Let $T \leq 10$ be the length of the episode

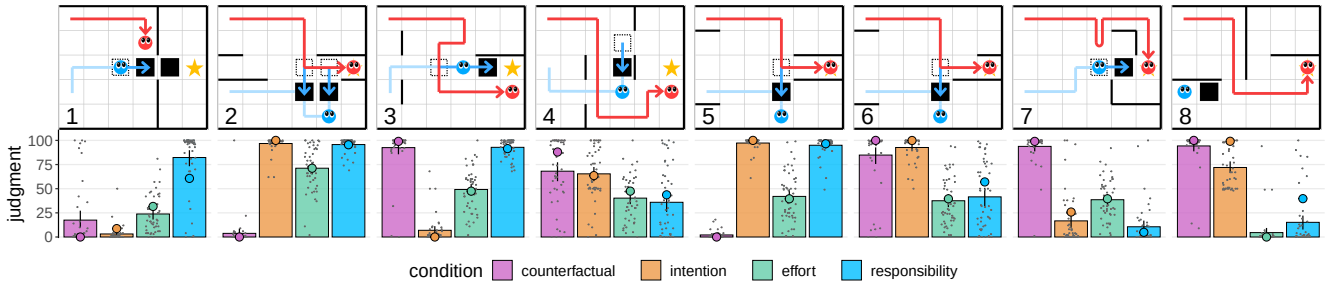

Figure 2: **Experiment 1 results**. Participants' judgments separated by condition (counterfactual, intention, effort, and responsibility) on a subset of trials. Note that the scale for intention judgments goes from "definitely hindering" being 0 and "definitely helping" being 100. Each trial diagram illustrates BLUE's path in light blue, box move actions with blue arrows, and RED's path. In all figures, bars show mean ratings, error bars are bootstrapped 95% confidence intervals, large points show model predictions, small points are individual judgments, RMSE = root mean squared error, and r = Pearson correlation coefficient.

and $H_{1,\ldots T}$ be the history of states, actions, and rewards that occurred. First, we can obtain an inference about BLUE's intentions using the distribution $p(\chi_{\mathrm{B}}^{T})$. Secondly, we can simulate what would have happened in a counterfactual scenario. We focus on whether RED would have succeeded if BLUE hadn't been there (e.g. Figure 1C). The model solves the Social MDP M_{R}^{0} or M_{R}^{2} without BLUE in the environment, but preserves any transitions that also occurred in $H_{1,\ldots T}$. Here, the environment is deterministic so all transitions result from agents' actions, but in stochastic environments, conditioning on H is crucial in distinguishing counterfactual from hypothetical simulation (Gerstenberg, 2022; Wu et al., 2022). Counterfactual episodes are run stochastically to capture uncertainty about RED's behavior: RED has a small chance $p = 0.1$ of stalling on each timestep. The model runs 1000 noisy counterfactual simulations to get a prediction for how likely RED would have succeeded if BLUE hadn't been there.

Responsibility for BLUE In line with prior research, we propose that responsibility judgments towards BLUE are driven by two factors: counterfactual judgments that reflect BLUE's causal role in the outcome, and inferences about BLUE's intentions. We coded both components to account for the outcome (i.e. used either raw or flipped values). Both components are then fit through a linear regression,

$$\text{responsibility}_{\mathrm{B}} = \alpha + \beta_1 \cdot \text{counterfactual} + \beta_2 \cdot \text{intention}.$$

Responsibility for RED We propose that responsibility for RED is inversely related to responsibility for BLUE: when BLUE is attributed much responsibility for the outcome, the amount assigned to RED is reduced, and vice versa. Formally,

$$\text{responsibility}_{\mathrm{R}} = \alpha - \beta \cdot \text{responsibility}_{\mathrm{B}}.$$

Alternative models

Effort is another factor that can be observed or inferred about an agent from their actions. The amount of effort a rational agent exerts reflects their desire for a particular outcome, which affects moral evaluations about them (Jara-Ettinger et al., 2016; Bigman & Tamir, 2016; Sosa et al., 2021). In Experiment 1, we test an alternative model of responsibility for BLUE that considers perceived effort in place of intentions, along with counterfactuals. We model BLUE's effort as a normalized sum of all action costs $c(a_{\mathrm{B}})$ incurred.

Another possible model of how people assign responsibility is that they rely on perceptual and physical cues instead of simulating counterfactuals and inferring mental states (Iliev et al., 2012; White, 2014). To test this, we construct a heuristic model that performs a linear regression over four perceptual features in each trial: the outcome, how many steps RED and BLUE each took, and how far any boxes were moved. In both experiments, we compare our responsibility model for BLUE to the heuristic model, as well as lesioned models that only include the counterfactual or intention component.

Experiment 1: Level-0 RED

In Experiment 1, we considered scenarios featuring a level-0 RED and a level-1 BLUE. Between conditions, participants were asked to make counterfactual, intention, effort, or responsibility judgments about what happened in each scenario. We tested how well the components of our model capture these judgments and predict responsibility overall.

Methods

All materials and data are available at: `https://github.com/cicl-stanford/counterfactual_agents`.

Participants The experiment was preregistered and posted on Prolific. 200 participants (*age*: M = 34, SD = 13; *gender*: 100 female, 88 male, 9 non-binary, 1 agender, and 2 undisclosed) were recruited and compensated $11/hour. They were randomly assigned to the *counterfactual*, *intention*, *effort*, or *responsibility* conditions with $n = 50$ in each.

Procedure Participants were introduced to the setting with RED and BLUE. They were guided through instructions with an example trial and then required to answer three comprehension questions correctly before proceeding to the main task. During the main task, they saw 24 different trials in

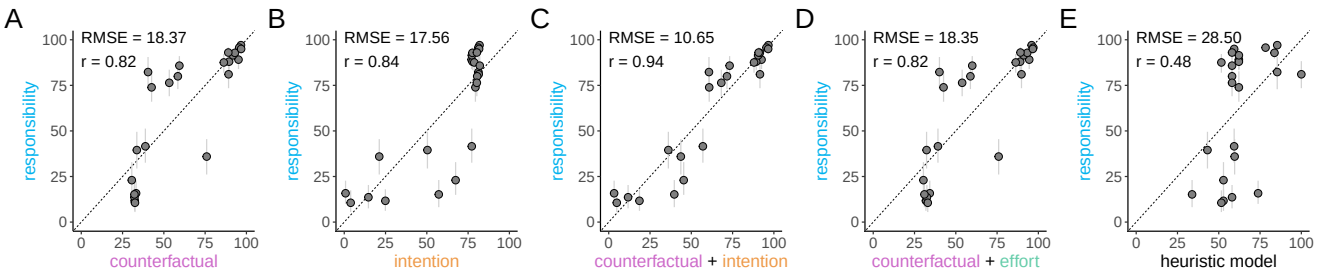

Figure 3: **Experiment 1 responsibility judgments**. Participants' mean judgments for BLUE compared to model predictions that consider (A) counterfactuals only, (B) intentions only, (C) both counterfactuals and intentions, (D) effort instead of intentions, and (E) a heuristic model. Model predictors use participants' mean judgments from the respective conditions.

a randomized order (see Figure 2 for some examples).

In each trial, participants watched what happened and then responded to a question with a video replay of the scenario available. In the *counterfactual* condition, participants were asked how much they agreed that "RED [would have / would still have] succeeded if BLUE hadn't been there." We used "would have" if the outcome was a failure and "would still have" if it was a success. Participants answered on a continuous slider from "not at all" (0) to "very much" (100). In the *intention* condition, participants were asked "What was BLUE intending to do?" and answered on a slider from "definitely hinder RED" (0) to "definitely help RED" (100) with the midpoint labeled "unsure" (50). In the *effort* condition, participants were asked "How much effort did BLUE exert?" with the slider endpoints labeled "very little" (0) and "very much" (100). Finally, the *responsibility* condition was similar to the counterfactual condition except that the question read "How responsible was BLUE for RED's [success/failure]?". The experiment took an average of 12 minutes (SD = 7) to complete.

Design Across the 24 trials in the experiment, we manipulated whether RED succeeded or failed (actual outcome), and whether RED would have succeeded or failed had BLUE not been there (counterfactual outcome). BLUE's intentions were also varied so that they could appear to be helping, hindering, or have ambiguous intentions. We manually generated BLUE's actions in each trial to create interesting interactions, but modeled them as a level-1 agent.

Results

Figure 2 shows participants' judgments in the different conditions across a subset of the 24 scenarios. Our model used a softmax of $\beta = 0.4$ and discount factor of $\gamma = 0.99$. The model captures much of the variance in participants' counterfactual judgments ($r = 0.95$, RMSE $= 14.73$), intention inferences ($r = 0.97$, RMSE $= 11.74$), and effort judgments ($r = 0.95$, RMSE $= 5.44$). To predict responsibility judgments, we fit five different Bayesian linear mixed effects models. The first considers counterfactuals only, the second considers intentions only, the third considers both counterfactuals and intentions (our model), the fourth considers counterfactuals and effort as an alternative, and the fifth is a heuristic model. All models included random intercepts and slopes for each participant. We used participants' mean counterfactual, intention, and effort judgments as predictors in the models.

Figure 3 shows that our model ('counterfactual + intention') qualitatively captures responsibility judgments well. It also has the highest correlation and lowest RMSE. For a quantitative comparison, we used approximate leave-one-out cross-validation, which takes into account the varying model complexity. Table 1 shows that our model performs best overall, and best fits the most individual participants' judgments (using the same cross-validation procedure).

Discussion

In this experiment, we found that participants' responsibility judgments for BLUE were well predicted by our model, which explains responsibility as a combination of counterfactual judgments about what would have happened had BLUE not been there (reflecting BLUE's causal role), and inferences about BLUE's intentions (reflecting BLUE's character). Neither component alone predicts responsibility as well, and intention is a better predictor than effort. For example, in trial 1, BLUE pushed a box in RED's way despite another box already blocking the star. Participants judged BLUE's causal role to be low (RED would have been unlikely to succeed even if BLUE hadn't been there), and BLUE's effort to be low, but still held

Table 1: **Experiment 1 model comparison**. "Δelpd" shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models, along with associated standard error. Lower numbers represent worse performance (Vehtari et al., 2017). "*n* best" is the number of individual participant judgments best fit by each model.

Model	Δelpd (se)	*n* best
counterfactual + intention	0 (0)	38
intention	−162.0 (18.5)	2
counterfactual	−178.6 (21.8)	6
counterfactual + effort	−179.1 (21.9)	4
heuristic	−445.2 (28.1)	0

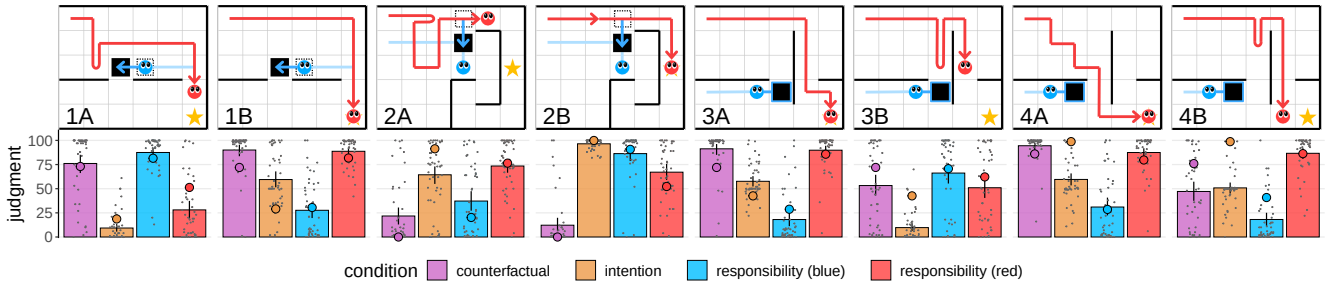

Figure 4: **Experiment 2 results**. Participants' judgments separated by condition (counterfactual, intention, and responsibility for BLUE and RED) on a subset of trials. Each pair features an (A) level-0 RED and (B) level-2 RED on the same scenario.

BLUE very responsible because of the strong inference that they intended to hinder. Conversely, in trials 5 and 6, participants inferred similar intentions and effort, but different counterfactuals, which set apart the respective responsibility judgments. In trial 5, RED could not have succeeded without BLUE's help, but in trial 6, the missing wall would have made that possible. In the next experiment, we explore more complex interactions involving reasoning beyond level-1, and investigate responsibility judgments for both agents.

Experiment 2: Level-2 RED

In Experiment 2, we introduced a level-2 RED. This agent plans relative to a level-1 BLUE's social goal and can, for example, wait for BLUE to move a box out of their way after correctly inferring BLUE's intention to help. Then, a level-3 BLUE, who reasons about a level-2 RED, would be able to act so as to *deceive* RED by deliberately signaling false intentions (knowing that RED assumes them to be level-1). For example, in trial 3B in Figure 4, BLUE moved right to pick up the box, and RED moved down, anticipating that BLUE would helpfully pull the box aside. However, BLUE didn't actually move the box, forcing RED to backtrack and ultimately fail. In that sense, higher-level reasoning enables difference-making beyond just the physical world: it becomes possible for an agent to play a causal role in the outcome by affecting the mental states of others. In this experiment, we extended our responsibility model to a wider range of social interactions, including scenarios in which BLUE makes no changes to the physical environment, but nevertheless affects RED's actions.

Participants were asked to make counterfactual, intention, and responsibility judgments about each scenario. We asked about responsibility for RED in addition to BLUE, and also asked some participants to give open-ended explanations about why RED succeeded or failed in each trial.

Methods

Participants The experiment was preregistered and posted on Prolific. 200 participants (*age*: M = 36, SD = 12; *gender*: 98 female, 93 male, 7 non-binary, 1 transgender, 1 undisclosed) were recruited and compensated $12/hour. They were randomly assigned to the *counterfactual*, *intention*, *responsibility*, or *explanation* conditions with $n = 50$ in each.

Procedure & Design The procedure and design were similar to that of Experiment 1. The *counterfactual* and *intention* conditions were identical. In the *responsibility* condition, participants were asked "How responsible was RED for the [success/failure]?" and "How responsible was BLUE for the [success/failure]?" on separate sliders from "not at all" (0) to "very much" (100). In the *explanation* condition, they were asked "Why did RED [succeed/fail]?" and typed their answers in a free-form text box. The experiment took an average of 21 minutes (SD = 12) to complete. We designed 24 trials consisting of 12 pairs in which the environment and BLUE 's actions were the same, but RED was either level-0 or level-2. Across the 12 pairs of trials, we manipulated the outcome and BLUE's intentions as in Experiment 1.

Results

Figure 4 shows participants' judgments in the different conditions across the 12 pairs of scenarios. Our model used a softmax of $\beta = 0.3$ and discount factor of $\gamma = 0.99$. The model accounted well for counterfactual judgments ($r = 0.8$, RMSE $= 23.05$) and intention inferences ($r = 0.72$, RMSE $= 27.81$), although some variance remains unexplained.

Responsibility We tested the same set of Bayesian models from Experiment 1 on responsibility judgments for BLUE, except for the 'counterfactual + effort' model. Because that model did not perform as well, we dropped the effort condition here. We again used participants' mean judgments as predictors in the models. The results are similar to those from Experiment 1. Figure 5 and Table 2 show that our model ('counterfactual + intention') best captures responsibility for BLUE qualitatively and quantitatively, for both overall and in-

Table 2: **Experiment 2 model comparison**. See Table 1 for column definitions. Lower Δelpd is worse in cross-validation.

Model	Δelpd (se)	*n* best
counterfactual + intention	0 (0)	26
intention	−92.1 (13.8)	15
counterfactual	−142.1 (17.5)	7
heuristic	−350.5 (25.4)	2

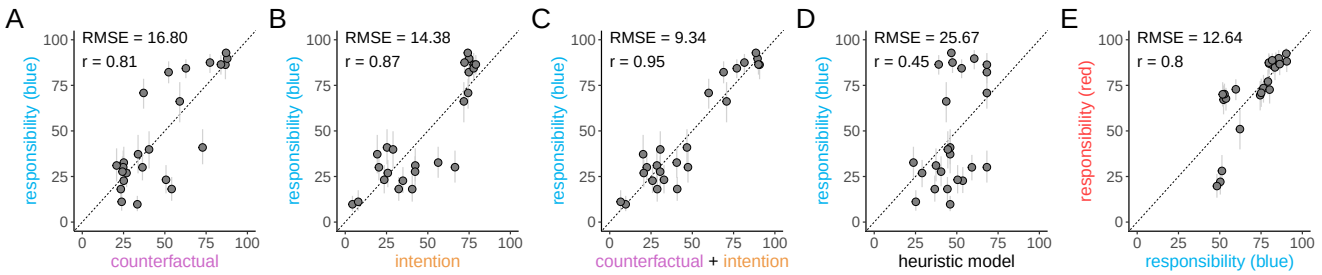

Figure 5: **Experiment 2 responsibility judgments**. Participants' mean judgments compared to model predictions considering (A) counterfactuals only, (B) intentions only, (C) both counterfactuals and intentions, and (D) a heuristic model. Plot (E) shows responsibility judgments for RED predicted as a function of corresponding responsibility judgments for BLUE.

dividual participant judgments. Figure 5E shows our responsibility model for RED, using corresponding judgments for BLUE as a predictor with a random intercept and slopes for each participant. This model performs decently overall.

Explanations Participants' open-ended explanations about why RED succeeded or failed in each trial were coded based on whether they mentioned the following features: the box, time, RED's actions, RED's mental states or actions requiring mentalizing, BLUE's actions, and BLUE's mental states or mentalizing actions. Overall, participants mentioned RED's actions more than RED's mental states, and vice versa for BLUE (Figure 6A). In both trials 3B and 4B, BLUE made no changes to the physical environment, but participants attributed different amounts of responsibility (Figure 5), which mapped onto different types of explanations (Figure 6B). In trial 3B, many participants noted BLUE's deception (e.g. "[BLUE] tricked [RED] into thinking she was going to move the box to help her, but once [RED] was stuck on that side of the wall, [BLUE] left the box where it was."). In contrast, in trial 4B, most participants faulted RED's own behavior and barely mentioned BLUE (e.g. "[RED] questioned their route and reversed, thus not having enough steps to reach the star.").

Discussion

Experiment 2 expanded on Experiment 1 by testing scenarios involving higher-level reasoning, modeling responsibility for RED in addition to BLUE, and analyzing open-ended explanations about the outcome. Like in Experiment 1, we found that responsibility for BLUE was best explained by a combination of beliefs about BLUE's causal role in the situation and their helping or hindering intentions. In turn, this predicted responsibility for RED well, although other factors are likely also at play. While RED's intentions are not apparent (as they have no social goal), it is still possible to construct counterfactuals in which they had acted differently. For example, in trial 1A, RED might be held responsible for the failure on the basis that, had they been level-2 instead of level-0, they would have succeeded like in trial 1B, where they waited for BLUE to move the box out of the way.

Our model captures participants' counterfactual simulations and intention inferences to a good extent, but there may be additional sources of uncertainty in people's judgments that are not yet accounted for. We found mixed results regarding situations in which BLUE seemingly affected RED's mental states. In both trials 3B and 4B, participants were uncertain about whether RED would have succeeded without BLUE, but they recognized BLUE's deceptive intentions more strongly in trial 3B, which drove responsibility judgments up. Perhaps they were less confident in trial 4B because RED backtracked less, or because of an asymmetry between positive and negative signaling. Future work is needed to resolve these cases of higher-level reasoning.

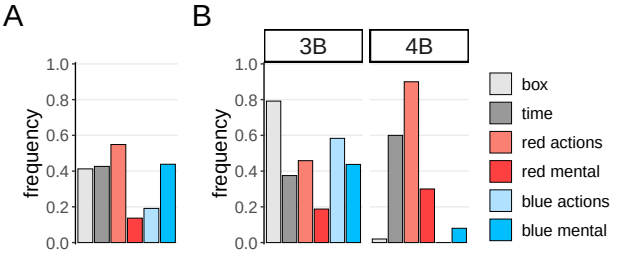

Figure 6: **Experiment 2 explanations**. Frequency of features mentioned (A) across all trials for a random subset of participants, and (B) specifically on trials 3B and 4B.

Conclusion

In this paper, we developed and tested a computational model of responsibility judgments that bridges mechanisms of counterfactual simulation and intention inference using a shared underlying generative planner. The planner captures people's intuitive theory of mind about agents' behavior. Across a variety of animated scenarios, our model captured participants' counterfactual simulations and intention inferences. Together, these two components predicted responsibility judgments better than alternative models of effort, heuristics, or either component alone. This model brings us closer to a formal, comprehensive understanding of how people attribute responsibility.

Acknowledgments

This work was supported by an NSF Graduate Research Fellowship (GRFP) to Sarah A. Wu, a grant from Stanford Symbolic Systems to Shruti Sridhar, and a grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI) to Tobias Gerstenberg. All experiments were approved by Stanford's Institutional Review Board.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654.

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.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

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

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(6), 936–975.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.), *Oxford Handbook of Causal Reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122-141.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition*, *40*(8), 1387–1401.

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

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 785.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128).

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021). Phase: Physically-grounded abstract social events for machine social perception. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 845–853).

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

Shu, T., Kryven, M., Ullman, T. D., & Tenenbaum, J. B. (2020). Adventures in Flatland: Perceiving social interactions under physical dynamics. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Sosa, F. A., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890.

Tejwani, R., Kuo, Y.-L., Shu, T., Katz, B., & Barbu, A. (2021). Social Interactions as Recursive MDPs. In *Proceedings of the 5th Conference on Robot Learning* (Vol. 164, pp. 949–958). PMLR.

Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (Vol. 22, pp. 1874–1882).

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.

White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science*, *38*(1), 38–75.

Wright, J. (2010). Beyond equilibrium: predicting human behaviour in normal form games. In *Proceedings of the Behavioral and Quantitative Game Theory on Conference on Future Directions.*

Wu, S. A., & Gerstenberg, T. (2023). If not me, then who? Responsibility and replacement. *PsyArXiv*.

Wu, S. A., Sridhar, S., & Gerstenberg, T. (2022). That was close! A counterfactual simulation model of causal judgments about decisions. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society* (pp. 3703–3710).

Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online Bayesian Goal Inference for Boundedly Rational Planning Agents. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 19238–19250).