Helping and hindering as counterfactual difference-making

Sarah A. Wu[1], Shruti Sridhar[2], and Tobias Gerstenberg[1]
[1]Department of Psychology, Stanford University
[2]Department of Computer Science, Stanford University

Abstract

What does it mean for one person to help or hinder another? We argue that there are at least two important factors: one, the persons's intention, and two, what difference their action made. While prior work has focused on inferring helping or hindering intentions, we show here that counterfactual simulation is critical for understanding whether someone actually helped or hindered. We develop and test a computational model of responsibility judgments for helping and hindering interactions that integrates intention inference (via inverse planning) with causal attribution (via counterfactual simulation). Experiment 1 investigates scenarios where one agent helps or hinders another by acting on the physical environment. Experiment 2 features more complex scenarios where one agent influences the other by signaling their intentions, sometimes without acting on the physical environment. Finally, Experiment 3 investigates how the agent who was helped or hindered is held responsible depending on how they updated their beliefs about the other from a previous interaction. Across all three experiments, responsibility judgments were best captured by a model that combines intention inferences with causal attributions.

Keywords: responsibility; intentions; causal reasoning; counterfactual simulation; theory of mind.

Corresponding author: Sarah A. Wu (sarahawu@stanford.edu). All code, experiments, data, analyses, and links to preregistrations are available at: `https://github.com/cicl-stanford/helping_hindering`.

Introduction

Few moral ideals are as deeply ingrained as that of the Good Samaritan: someone who kindly helps a stranger in need. Good Samaritans are commonly depicted in popular culture, held up as role models for children, and even codified into law (e.g. *California Health & Safety Code § 1799.102*, 2009). These images endure because they reflect the fact that helping is a cornerstone of human social life. Our prosociality is widespread and near-universal (Aknin et al., 2013; Henrich et al., 2005; House et al., 2013; Rossi et al., 2023), sets us apart from other species (Drayton & Santos, 2016; Warneken & Tomasello, 2009), and emerges early in life (Dahl, 2015; Warneken & Tomasello, 2006; Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman, 1992). While a large body of research has documented the diverse ways in which people help one another, and illuminated when and why people help (for reviews, see e.g. Eisenberg, Spinrad, & Knafo-Noam, 2015; Penner, Dovidio, Piliavin, & Schroeder, 2005), less is understood about a more basic question. How do people represent the concept of helping and recognize when it happens? On the flip side, how do people recognize acts of hindering or harming?

One influential theory formalizes helping and hindering in terms of utility adoption (Powell, 2022; Schlingloff-Nemecz et al., 2025). This approach builds on a formal model of action understanding that assumes that people are rational utility-maximizers (Gergely & Csibra, 2003; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Jara-Ettinger, Schulz, & Tenenbaum, 2020). Accordingly, people act in ways that generally maximize their expected rewards and minimize their expected costs. Social behavior arises when one agent incorporates another agent's utility into their own (Powell, 2022). For example, an agent intends to help another agent if they choose actions that increase the other's expected utility, sometimes even at a cost to themselves. In contrast, the agent intends to *hinder* the other agent if they act to decrease the other's expected utility.

This utility-based approach to action understanding not only predicts how someone will act given their goals and intentions, but also captures what inferences one can make about an agent's mental states from their actions. Such mental state inferences can be modeled computationally as Bayesian inverse planning (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Jern, Lucas, & Kemp, 2017). Roughly, we infer that an agent must have a goal such that their observed behavior is a sensible way of achieving it. Inverse planning models accurately capture people's inferences of helping and hindering intentions (Baker, Goodman, & Tenenbaum, 2008; Chandra, Li, Tenenbaum, & Ragan-Kelley, 2023; Netanyahu, Shu, Katz, Barbu, & Tenenbaum, 2021; Shu, Kryven, Ullman, & Tenenbaum, 2020; Ullman et al., 2009).

However, intending to help is not the same as actually helping. A young child may intend to help with grocery shopping, but ultimately slow things down compared to their parent shopping alone. While inverse planning can recover the child's intentions to help, it does not track whether those intentions were successfully realized. Determining whether the child actually helped is fundamentally a causal question, and requires contrasting what happened with what would have happened had the child not been there. Such counterfactual comparisons are key to defining what it means to help, hinder, or harm across many domains. In philosophy, the *counterfactual comparative account* of harm defines harm as causing someone to be worse off than they would have been otherwise (Feinberg, 1986;

Klocksiem, 2012; Purves, 2019). In computer science, researchers have argued that helping and harming are inherently causal notions requiring counterfactual analysis (Beckers, Chockler, & Halpern, 2024; Richens, Beard, & Thompson, 2022).[1] And in the law, criminal liability has long required proving not only *mens rea* (a guilty mind), but also *actus reus* (a guilty act), distinguishing the intent to harm from the act of harm (Lagnado & Gerstenberg, 2017). Yet, while counterfactual accounts of help and harm have been proposed across these fields, they have not been tested as a cognitive theory of how people actually understand these concepts.

Counterfactual thinking plays a central role in causal reasoning. For example, in the physical domain, when asked whether one ball caused another to go through a goal, people mentally simulate how things would have turned out if the first ball had not been there (Beller & Gerstenberg, 2025; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Gerstenberg & Stephan, 2021; Zhou, Smith, Tenenbaum, & Gerstenberg, 2023). A computational model that generates counterfactual scenarios using an intuitive physics engine (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg & Tenenbaum, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017; but see Ludwin-Peery, Bramley, Davis, & Gureckis, 2021) accurately captured participants' causal judgments. Participants' causal judgments increased the more certain they were that the counterfactual outcome would have been different.

In the social domain, counterfactual simulation has been argued to be important, too (Cushman, 2024; Gerstenberg, 2024; Lipe, 1991). For example, to determine what causal role a person played in bringing about an outcome, one may consider what would have happened without them, or if they had acted differently (Chockler & Halpern, 2004; Halpern & Pearl, 2005; Lagnado, Gerstenberg, & Zultan, 2013; Lewis, 1973; Pearl, 2000; Wu & Gerstenberg, 2024). The more likely the outcome would have been different had a person not acted as they did, the more responsible participants tended to hold them. Thoughts about how things could have gone differently also feature centrally in theories of how people explain behavior (Byrne, 2016; Kahneman & Tversky, 1982; Lombrozo, 2010; Macrae, Milne, & Griffiths, 1993; Markman, Gavanski, Sherman, & McMullen, 1993; Petrocelli, Percy, Sherman, & Tormala, 2011). However, despite the theoretical parallels between physical and social causation, no work has directly tested what role counterfactual simulation plays for determining whether one agent helped or hindered another.

In this paper, we develop a computational model of responsibility judgments for helping and hindering scenarios that goes beyond inferring an agent's intentions and incorporates whether the agent's actions made a difference to the outcome. The responsibility model has two components. First, it infers an agent's intentions from its actions by inverting a model of the agent's planning process. Second, it evaluates an agent's causal role by simulating how things would have unfolded if the agent hadn't been there. The integration of intentions and causation is consistent with prior theories of responsibility that highlight the influence of person inferences and causal attributions (Gerstenberg et al., 2018; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Malle, Guglielmo, & Monroe, 2014; Sartorio, 2007; Shaver, 1985; Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021; Weiner, 1995). Here, we go beyond this earlier work by formally instantiating both

[1] Our experiments focus on hindering (i.e. obstructing progress) rather than harming (i.e. causing damage or injury), but we believe that counterfactual comparisons are critical for both.

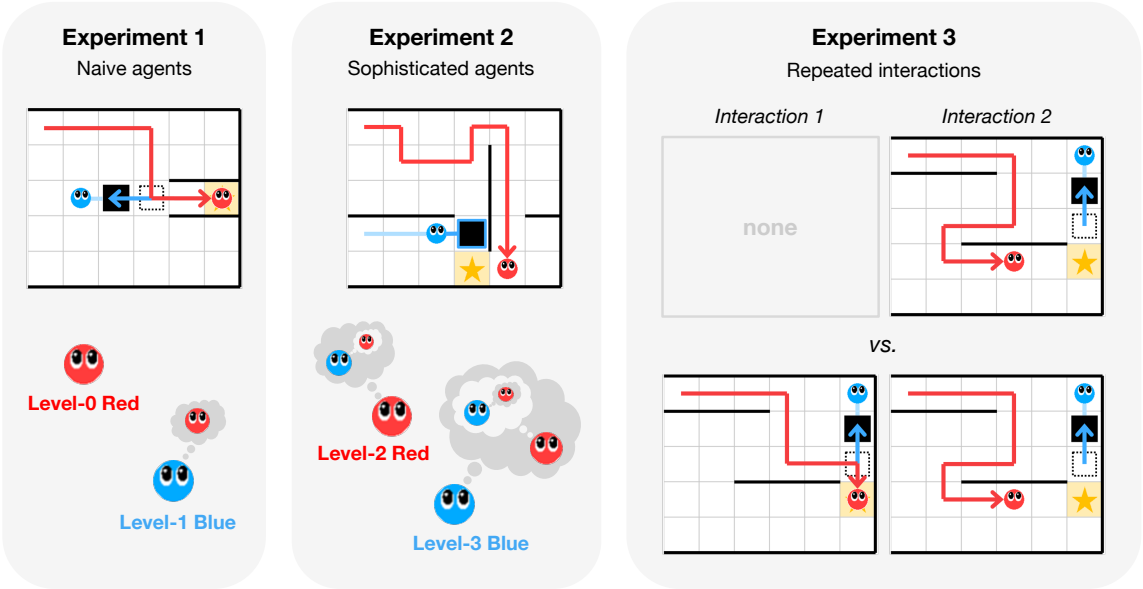

Figure 1. **Overview of experiments**. In Experiment 1, we investigated interactions between a naive RED, who plans only towards the star, and a naive BLUE, who plans to help or hinder RED. In Experiment 2, we investigated a sophisticated RED, who additionally infers BLUE's intentions, and a sophisticated BLUE, who reasons about a sophisticated RED. In Experiment 3, we studied repeated interactions and the effect of prior beliefs about others on responsibility judgments.

components as computations that operate over a shared underlying representation.

To test the model, we designed a dynamic environment where one agent (RED) tries to reach a star within a limited period of time, and another agent (BLUE) intends to either help or hinder RED (Figure 1). BLUE can move boxes, but RED can't. For example, in Figure 1A, RED started in the top left corner, BLUE pulled a box out of RED's way, and RED succeeded in reaching the star. The agents take turns moving, and BLUE can only move a box once in each scenario. We model agents at different levels of sophistication (Camerer, Ho, & Chong, 2004; Wright, 2010). A *naive* RED agent only reasons about the physical world, while a *sophisticated* RED infers and plans around BLUE's intentions. This agent can, for example, try to avoid BLUE or wait for BLUE to move a box out of the way. A naive BLUE agent plans to help or hinder a naive RED, and a sophisticated BLUE plans to help or hinder a sophisticated RED. Because the sophisticated BLUE is aware that the sophisticated RED maintains beliefs about it (assuming it to be naive), BLUE can act in ways so as to influence those beliefs, such as appearing to help when in fact it intends to hinder (Figure 1B).

Across three experiments, we test how well the model captures participants' judgments about the agents' intentions, counterfactual outcomes, and responsibility. In Experiments 1 and 2, we study how people attribute responsibility to the BLUE agent. Experiment 1 features naive agents only, and Experiment 2 features sophisticated agents who exhibit more complex behaviors. In Experiment 3, we investigate responsibility judgments for both BLUE and RED, and specifically look at how a sophisticated RED agent may be held

responsible for the outcome on the basis of its *beliefs* about BLUE's intentions.

Computational model

We represent our setting as a set of Markov Decision Processes (MDPs), which capture how social interactions between rational agents unfold over time in a dynamic world (Baker et al., 2017; Jara-Ettinger, 2019). We extend the MDPs to allow agents to have social goals and presentational goals in addition to physical goals (Btesh, Lagnado, & Gerstenberg, 2025; Tejwani, Kuo, Shu, Katz, & Barbu, 2021; Tejwani et al., 2022). The MDP for agent $i \in \{\text{RED}, \text{BLUE}\}$ at level ℓ is the tuple

$$M_i^\ell = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \chi_i, \rho_i, g_i, R_i^\ell, \gamma, H \rangle, \tag{1}$$

where $\mathcal{S}$ is the space of all states s; $\mathcal{A}$ is the joint action space of all pairs of actions a_{RED}, a_{BLUE}; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ dictates the transition probabilities; $\chi_i \in X$ is agent i's social goal; $\rho_i \in P$ is agent i's presentational goal; $g_i \in G$ is agent i's physical goal; $R_i^\ell : \mathcal{S} \times \mathcal{A}_i \times X \times P \times G \to \mathbb{R}$ is agent i's reward function at level ℓ; $\gamma \in (0, 1)$ is a reward discount factor, and H is a finite time horizon. In total, we construct four MDPs: M^0_{RED} for the naive RED, M^1_{BLUE} for the naive BLUE, M^2_{RED} for the sophisticated RED, and M^3_{BLUE} for the sophisticated BLUE.

Planning

Each agent's reward function depends on the agent's actions and goals. The reward discount γ encourages agents to reach their goals sooner rather than later. The RED agent only has a physical goal g_{RED}, which is to reach the star, so $\chi_{\text{RED}} = \rho_{\text{RED}} = 0$. If time runs out, then RED gets partial reward based on its navigable distance from the goal (ignoring boxes), which incentivizes it to get as close as possible. The naive and sophisticated RED reward functions are the same; the difference between the two agents is that the naive RED only reasons about the physical world, while the sophisticated RED infers BLUE's intentions and anticipates BLUE's next possible action at each time step.

BLUE has no physical goal, so $g_{\text{BLUE}} = 0$. Its social goal χ_{BLUE} is a constant that scales RED's reward into some of its own reward such that when χ_{BLUE} is positive, BLUE attempts to help RED, and when χ_{BLUE} is negative, BLUE attempts to hinder RED. BLUE's presentational goal ρ_{BLUE} is a constant that incorporates RED's *beliefs* about BLUE's actual social goals into some of BLUE's reward. When ρ_{BLUE} is positive, BLUE wants to *appear* helpful by maximizing RED's estimate of the value of χ_{BLUE}, while the opposite is true when ρ_{BLUE} is negative. Rewards are computed and experienced at each timestep, so BLUE aims to not only achieve its presentational goal but also maintain it for as long as possible. The naive BLUE agent does not have a presentational goal because it assumes a naive RED who only reasons about the world. A sophisticated BLUE agent can have distinct social and presentational goals. For example, it may want to hinder ($\chi_{\text{BLUE}} < 0$) but appear helpful ($\rho_{\text{BLUE}} > 0$).

Agents recursively solve the MDPs at each level of sophistication to estimate each other's rewards. Each MDP M_i^ℓ is solved by computing a state-value function $Q_i^\ell(s, a, \cdot)$, which captures the expected reward the agent would receive if it were to take action a while in state s. Computing the Q-function requires knowing the other agent's goals and beliefs.

However, RED does not know BLUE's social goal χ_{BLUE} and the sophisticated BLUE does not have access to RED's beliefs about itself. Thus, the agents infer these latent states from observed actions at each timestep t through a Bayesian update:

$$p(\widetilde{\chi}^{t}_{\text{BLUE}} \mid s^{t-1},\ a^{t-1}_{\text{BLUE}}) \propto p(a^{t-1}_{\text{BLUE}} \mid s^{t-1},\ \widetilde{\chi}^{t-1}_{\text{BLUE}})\ p(\widetilde{\chi}^{t-1}_{\text{BLUE}}) \tag{2}$$

starting from a uniform prior at $t = 0$.

The Q-functions for each agent's MDP can then be used to generate an action plan, or policy, that maximizes the agent's total expected reward. Both agents model the other as having a *softmax* policy where actions are selected based on the Q-function:

$$p(a \mid s, \cdot) \propto \exp\Big(\beta \cdot \widetilde{Q}(s, a, \cdot)\Big). \tag{3}$$

The inverse temperature parameter β captures the agent's level of "randomness" while acting, and maintains uncertainty over each agent's estimate of the other. When choosing their own actions, agents use *argmax* policies where they choose the highest Q-valued action from each possible state, starting from their initial locations. In the case of ties, agents prefer to stay in place over any move action. If multiple move actions are tied for the highest Q-value, then the agent selects uniformly at random among them. We assume that these solutions approximate people's intuitive theories of how agents might interact based on their mental states, capacities, and situational constraints (Baker et al., 2017, 2009; Jara-Ettinger et al., 2016; Shu et al., 2020; Ullman et al., 2009). To predict responsibility in each scenario, we use MDPs to infer intentions and attribute causation (Figure 2).

Intention inference

To infer BLUE's intentions the model computes the final posterior on BLUE's social goal χ_{BLUE} at the end of each trial, which takes into account its full sequence of actions over the H timesteps:

$$\text{Intention} = p(\widetilde{\chi}^{H}_{\text{BLUE}}). \tag{4}$$

For the sophisticated BLUE, this distribution is marginalized over the presentational goal ρ_{BLUE}.

Causal attribution

We assume that people assess an agent's causal contribution by considering a relevant counterfactual situation and simulating what would have happened in their minds (Gerstenberg, 2024; Gerstenberg & Tenenbaum, 2017; Kahneman & Tversky, 1982). Here, we consider whether RED would have succeeded if BLUE hadn't been there. We simulate counterfactual scenarios by removing BLUE from the MDP and then running RED's policy. Each scenario is conditioned on the history of states and actions that actually happened (Gerstenberg, 2022; Pearl, 2000; Wu, Sridhar, & Gerstenberg, 2022). The model conditions RED's counterfactual path on its actual path by repeating the actual path with probability p_{follow} at each timestep. Once the counterfactual path deviates from the actual path, RED follows its policy in the counterfactual MDP without BLUE for the remaining timesteps. More precisely, the model computes the likelihood that RED's counterfactual

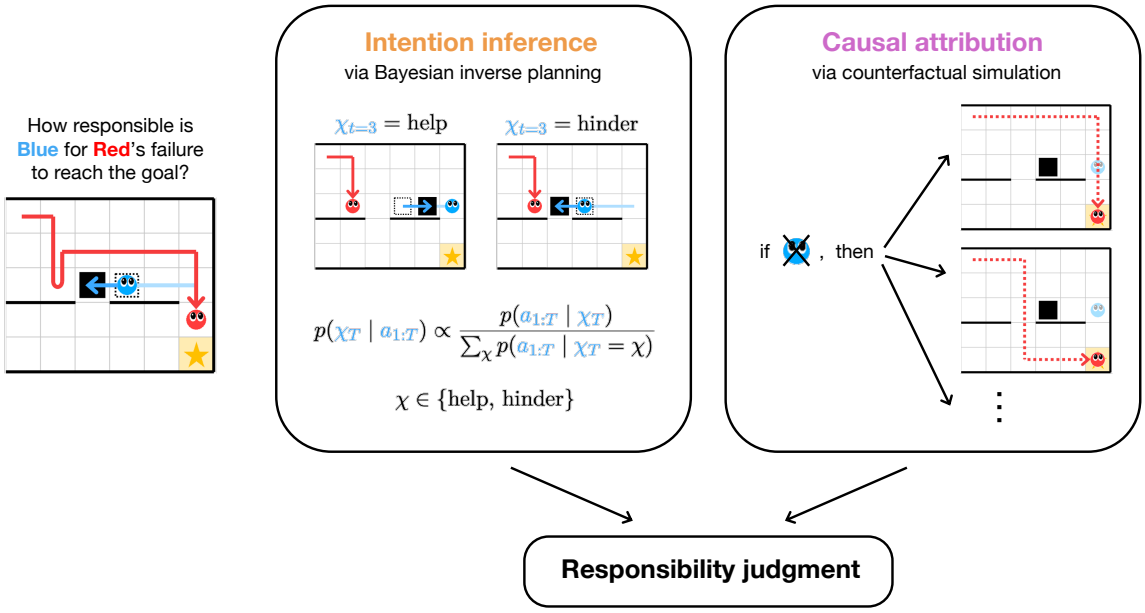

Figure 2. **Computational model of responsibility**. In this scenario, RED started in the top left corner. BLUE pushed a box into its path, causing RED to backtrack and miss the goal in time. The model predicts how responsible BLUE is for the outcome through two processes (1) *intention inference*: inferring the relative likelihood of BLUE's intention to help or hinder through Bayesian inverse planning, conditioning on BLUE's observed action at each timestep, and (2) *causal attribution*: running noisy counterfactual simulations about what would have happened if BLUE hadn't been there. The model combines these two processes to predict responsibility judgments.

outcome $o' \in \{\text{success}, \text{fail}\}$ without BLUE would have been different from what actually happened o:

$$\text{Counterfactual} = p(o' \neq o \mid M_{\text{RED}}). \tag{5}$$

RED's policy in each simulation is softmax rather than *argmax* to introduce uncertainty about RED's counterfactual behavior, controlled via an inverse temperature parameter (Equation 3). We aggregate many noisy simulations for each scenario to get a graded prediction for how likely RED would have succeeded without BLUE.

Responsibility from intention inference and causal attribution

We propose that people hold an agent more responsible the more certain they are that (1) the agent intended to bring about the outcome, and that (2) the agent's actions made a difference to the outcome (Figure 2). To predict responsibility judgments, we fit a linear combination of intention inferences and counterfactual judgments:

$$\text{Responsibility} = \alpha + \beta_1 \cdot \text{Intention} + \beta_2 \cdot \text{Counterfactual},$$

where α is an intercept and β_1 and β_2 capture the relative influence of each component.

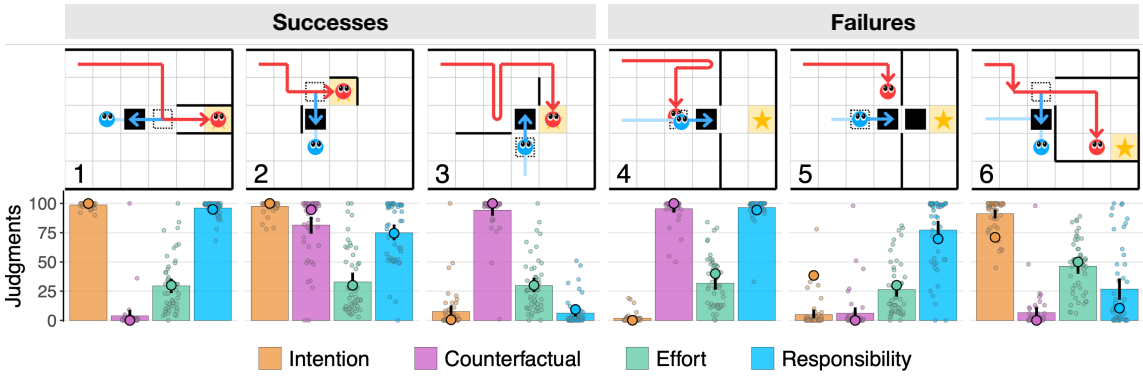

Figure 3. **Select trials from Experiment 1**. Participants' judgments separated by condition (intention, counterfactual, effort, and responsibility) on a subset of trials from Experiment 1. Note that the scale for intention goes from "definitely hindering" being 0 to "definitely helping" being 100. Each trial diagram illustrates BLUE's path in light blue, box movements with blue arrows, and RED's path with red arrows. Bars show mean ratings, error bars are bootstrapped 95% confidence intervals, large points show model predictions, and small points are individual judgments.

Alternative models

We compare the responsibility model to lesioned models that only include the intention or counterfactual component, as well as alternative models, such as one that considers the amount of effort an agent exerted (Jara-Ettinger et al., 2016). Prior work has shown that the more effort an agent exerted to bring about a negative outcome, the more negatively they tended to be judged (Bigman & Tamir, 2016; Sosa et al., 2021). We model effort as the normalized sum of all action costs incurred, and test a model that combines counterfactual judgments with effort inferences in Experiment 1.

It is also possible that people assign responsibility by relying on perceptual and kinematic features instead of inferring mental states and simulating counterfactuals (Cavallo, Koul, Ansuini, Capozzi, & Becchio, 2016; Iliev, Sachdeva, & Medin, 2012; McEllin, Sebanz, & Knoblich, 2018; White, 2014). Causal events and social interactions are sometimes perceived rapidly from motion cues alone (Heider & Simmel, 1944; McMahon & Isik, 2023; Michotte, 1963; Scholl & Tremoulet, 2000; Shu et al., 2020; Shu, Peng, Zhu, & Lu, 2021). We constructed a heuristic model that predicts responsibility judgments using a linear regression based on four features in each trial: the outcome, the number of steps that RED took, the number of steps that BLUE took, and the number of steps any boxes were moved. We chose these features because they are directly observable, related to each agent's behavior, and do not require mental simulation to compute.

Results

Experiment 1: Naive agents

In Experiment 1, participants watched scenarios showing interactions between naive RED and BLUE agents, and made judgments about what happened using sliders. Between

conditions they were either asked to judge what BLUE intended to do (from "definitely hinder" to "definitely help"), how likely RED would have succeeded if BLUE hadn't been there (from "not at all" to "very much"), how much effort BLUE exerted (from "very little" to "very much"), or how responsible BLUE was for the outcome (from "not at all" to "very much"). See the Methods section for the full materials.

Figure 3 shows participants' judgments for a subset of scenarios. In the first two scenarios, participants made similar judgments about how likely BLUE was helping and how much effort BLUE exerted. However, RED could not have succeeded without BLUE in scenario 1, whereas it could have in scenario 2. BLUE was held more responsible for the success in scenario 1 than in scenario 2. In scenario 3, RED succeeded *despite* BLUE's intention to hinder, so BLUE was attributed little responsibility for the success. In scenarios 4 and 5, BLUE intended to hinder by blocking RED's access to the goal. However, if BLUE hadn't been there, RED would have succeeded in scenario 4 but not 5, hence the difference in responsibility judgments. In scenario 6, BLUE's causal role was judged to be low, but BLUE actually intended to help so it was not held very responsible for the failure.

Overall, the individual model components capture much of the variance in participants' counterfactual judgments (Pearson $r = 0.97$, RMSE = 11.55) and intention inferences ($r = 0.93$, RMSE = 16.94). In addition, effort judgments were modeled well as the total ac-

Table 1

Model fits and comparison in Experiment 1. *Estimated posterior means and 95% credible intervals for different predictors in the models tested in Experiment 1. All models included random intercepts for participants. Credible effects are indicated with an asterisk* (meaning that the 95% credible interval excludes 0). r = Pearson correlation coefficient and RMSE = root mean squared error. "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, Gelman, & Gabry, 2017). 'n best' is the number of individual participants whose judgments were best predicted by each model.*

Model	Predictor	Estimate	r	RMSE	Δelpd (se)	n best
Intention + Counterfactual	Intercept	2.68 [−0.74, 6.15]	0.93	11.80	0 (0)	30
	Intention	0.69 [0.64, 0.73]*				
	Counterfactual	0.25 [0.21, 0.29]*				
Intention	Intercept	1.05 [−2.73, 4.68]	0.90	14.39	−72.0 (10.6)	13
	Intention	0.83 [0.79, 0.87]*				
Effort + Counterfactual	Intercept	59.11 [55.15, 63.04]*	0.74	21.78	−312.7 (26.0)	3
	Effort	−0.76 [−0.90, −0.62]*				
	Counterfactual	0.71 [0.66, 0.76]*				
Counterfactual	Intercept	42.80 [40.23, 45.39]*	0.68	23.62	−364.2 (30.5)	1
	Counterfactual	0.58 [0.53, 0.62]*				
Heuristic	Intercept	107.16 [100.34, 114.46]*	0.65	24.59	−403.8 (36.8)	3
	Outcome	6.05 [2.24, 9.69]*				
	RED steps	−8.37 [−9.18, −7.56]*				
	BLUE steps	9.68 [7.64, 11.71]*				
	Box moves	−3.12 [−9.09, 2.83]				

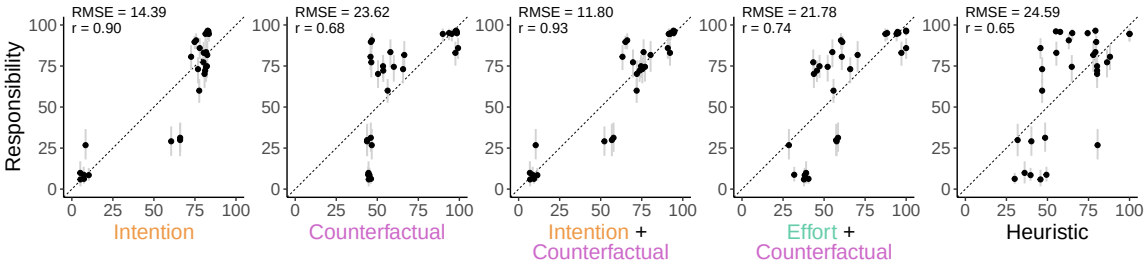

Figure 4. **Responsibility model predictions in Experiment 1**. Participants' mean judgments for BLUE compared to model predictions that consider intentions only, counterfactuals only, both intentions and counterfactuals, effort instead of intentions, with counterfactuals, and a heuristic model. Model predictors use participants' mean judgments from the respective conditions. Error bars are bootstrapped 95% confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

tion costs incurred by BLUE in each trial ($r = 0.95$, RMSE = 8.54). To predict responsibility judgments, we fit five Bayesian linear mixed effects models using different combinations of mean intention, counterfactual, and effort judgments as predictors (Figure 4 and Table 1). The 'intention + counterfactual' model predicts responsibility judgments best on all measures (Table 1). This model best fit overall judgments ($r = 0.93$, RMSE = 11.80) as well as the most individual participants (30 out of 50). Both the intention and counterfactual predictors were positive and credible. Notably, counterfactual judgments predict responsibility better when combined with intention judgments than with effort judgments. While responsibility can be sensitive to effort (Bigman & Tamir, 2016; Sosa et al., 2021), it additionally matters what that effort was directed towards. Intentions can distinguish between effort to bring about an outcome and effort to prevent that same outcome (e.g. scenarios 2 vs. 3 in Figure 3).

The results of Experiment 1 indicate that responsibility for helping and hindering is best explained by a combination of considering what BLUE was intending to do and how much of a difference BLUE's actions made to the outcome. We instantiated these two cognitive processes in the model using inverse planning to capture intention inferences, and counterfactual simulations to capture causal attributions. The more certain participants were that BLUE intended the outcome and played a critical causal role in bringing it about, the more responsible they held BLUE. So far, we have studied relatively simple settings where BLUE influences RED through direct interventions on the environment. In Experiment 2, we ask whether the model extends to richer social dynamics where sophisticated agents explicitly reason about each other's mental states, and where BLUE can influence RED by strategically manipulating its beliefs.

Experiment 2: Sophisticated agents

In Experiment 2, participants watched helping and hindering interactions between sophisticated agents. A sophisticated RED agent infers BLUE's intentions and can, for example, anticipate help by waiting for BLUE to move a box out of the way. This enables a sophisticated RED to succeed in scenarios where a naive RED wouldn't. A sophisticated

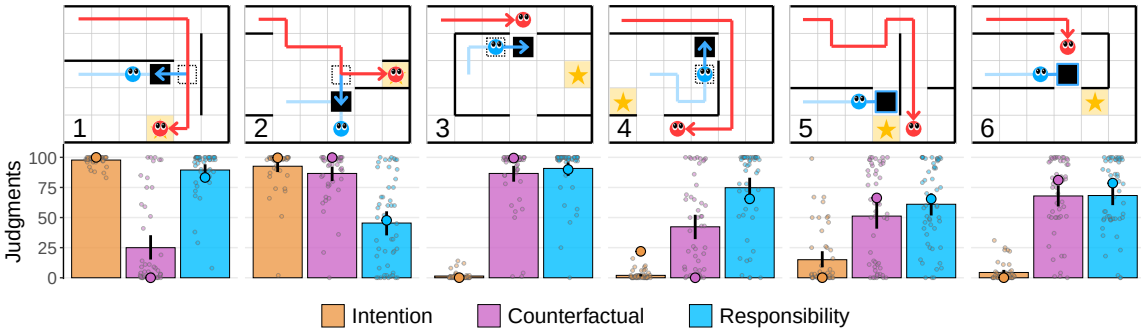

Figure 5. **Select trials from Experiment 2**. Participants' judgments separated by condition (intention, counterfactual, and responsibility) on a subset of trials from Experiment 2. Note that the scale for intention goes from "definitely hindering" being 0 and "definitely helping" being 100. Bars show mean ratings, error bars are bootstrapped 95% confidence intervals, large points show model predictions, and small points are individual judgments.

BLUE reasons about a sophisticated RED and maintains distinct social and presentational goals. Under certain combinations of these goals, BLUE can *deceive* RED by signaling false intentions (knowing that RED assumes it to be naive; see Oey, Schachner, & Vul, 2023; Schulz, Alon, Rosenschein, & Dayan, 2023). For example, in scenario 5 in Figure 5, BLUE initially appears to help by moving towards the box and picking it up. However, it then

Table 2

Model fits and comparison in Experiment 2. *Estimated posterior means and 95% credible intervals for different predictors in the models tested in Experiment 1. All models included random intercepts for participants. Credible effects are indicated with an asterisk*. r = Pearson correlation coefficient and RMSE = root mean squared error. "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. 'n best' is the number of individual participants whose judgments were best predicted by each model.*

Model	Predictor	Estimate	r	RMSE	Δelpd (se)	n best
Intention + Counterfactual	Intercept Intention Counterfactual	5.24 [0.48, 10.09]* 0.38 [0.30, 0.45]* 0.55 [0.46, 0.63]*	0.92	10.64	0 (0)	14
Counterfactual	Intercept Counterfactual	19.62 [15.55, 23.65]* 0.86 [0.79, 0.93]*	0.86	13.62	−46.5 (10.2)	21
Intention	Intercept Intention	4.58 [−0.30, 9.55] 0.71 [0.66, 0.77]*	0.83	14.90	−69.3 (12.8)	15
Heuristic	Intercept Outcome RED steps BLUE steps Box moves	67.29 [56.92, 77.09]* 16.72 [12.24, 21.39]* −3.69 [−4.58, −2.75]* −1.12 [−3.37, 0.94] 17.43 [11.51, 23.69]*	0.50	23.43	−243.1 (22.5)	0

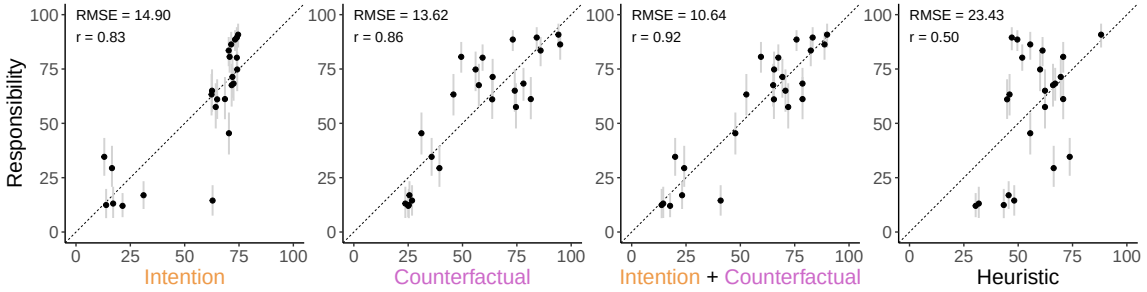

Figure 6. **Responsibility model predictions in Experiment 2**. Participants' mean judgments compared to model predictions considering intentions only, counterfactuals only, both intentions and counterfactuals, and a heuristic model. Model predictors use participants' mean intention and counterfactual judgments. Error bars are bootstrapped 95% confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

leaves the box where it is because BLUE actually intended to hinder, forcing RED to backtrack and ultimately fail to reach the goal. Participants recognized BLUE's actual intentions and often commented on its presentational goals (e.g. "Some rounds were easy to decide as 'intended to hinder' by how crafty Blue would lure Red into a false sense that [Red] would be helped, but then would get blocked at the right moment.").

Figure 5 illustrates participants' judgments across different outcomes and BLUE's intentions, from unequivocally helping (scenarios 1 and 2), to unequivocally hindering (scenarios 3 and 4), to hindering but appearing to help (scenarios 5 and 6). We dropped the effort condition here because it did not predict responsibility as well as intentions in Experiment 1. Overall, the individual model components captured much of the variance in participants' intention inferences ($r = 0.94$, RMSE = 15.56) and counterfactual judgments (Pearson $r = 0.88$, RMSE = 22.08). Using participants' intention and counterfactual judgments, we tested a similar set of Bayesian models as in Experiment 1 to predict responsibility judgments (Figure 6). Consistent with Experiment 1, the 'intention + counterfactual' model best explains overall responsibility judgments on all metrics, with $r = 0.92$, RMSE = 10.64, and best cross-validation performance (Table 2). Both the intention and counterfactual and predictors were positive and credible. However, individual participants showed more variation in model fits, with 20 out of 50 best described by the 'counterfactual' model, 16 best described by the 'intention' model, and 14 best described by the full model.

One possible explanation for the greater individual variation in model fits is that participants were more uncertain about both BLUE's intentions and the counterfactual outcomes in this experiment compared to Experiment 1. Exploratory Bayesian regression analyses showed that participants' responses were credibly closer to the midpoint of each scale in Experiment 2 than in Experiment 1. Intention judgments were closer by 2.08 points (95% CrI [1.27, 2.89], Bayes Factor $BF_{10} > 100$) and counterfactual judgments were closer by 6.72 points (95% CrI [5.93, 7.49], $BF_{10} > 100$). This increased uncertainty may reflect the greater complexity of the agents' mental models. Intentions were harder to read when BLUE's behavior was the product social *and* presentational goals, while counterfactual outcomes may have been harder to assess when the sophisticated RED actively coordinated

with BLUE rather than navigating straightforwardly toward the goal. As a result of this increased uncertainty, participants may have differed in which aspects they considered more heavily when assigning responsibility. Nonetheless, the overall results replicate and extend the findings of Experiment 1. Responsibility judgments were again sensitive to how certain participants were that BLUE made a causal difference to the outcome, but in these scenarios, that difference was not limited to physical changes in the environment. When agents plan recursively with each other in mind, it becomes possible for one agent to hinder another one by merely affecting its mental states.

So far, we have focused BLUE's responsibility by evaluating BLUE's intentions to help or hinder RED, and considering what would have happened if BLUE hadn't been there. In Experiment 3, we additionally explore how people attribute responsibility to RED. Prior work has shown that agents are blamed more for a negative outcome when they could have foreseen or prevented it (Lagnado & Channon, 2008; Malle et al., 2014), a pattern that contributes to victim blaming when victims are faulted for "not knowing better" about the perpetrator or the situation (Branscombe, Owen, Garstka, & Coleman, 1996; Niemi & Young, 2016). We investigate whether such judgments can be explained in our setting, by simulating counterfactuals of what would have happened if RED had held a different belief about BLUE's intentions.

Experiment 3: Hinder me once, blame on you; hinder me twice, blame on me

To manipulate RED's prior beliefs about whether BLUE will try to help or hinder, we designed trials where pairs of sophisticated RED and BLUE agents interact twice in the same environment. Some participants saw both rounds of interaction (*with prior* condition), while others only saw the second round (*no prior* condition). We further designed two types of trials, half in which BLUE attempted to help RED and the other half in which BLUE hindered RED. In helping trials, RED succeeded with BLUE's help in the first round, but then chose different actions and failed in the second round. For example, in Figure 7, BLUE pulled a box out of RED's way in both rounds A and B, but RED chose a different route and failed to reach the goal in round B. In hindering trials, RED failed after being hindered by BLUE in both rounds in the exact same manner. For example, in Figure 7, RED was forced to backtrack after its path was blocked by BLUE in both rounds. Note that in both types of trials, RED failed in the second round, for which participants were asked to infer BLUE's intention and assign responsibility to both agents.

With this trial design, the first round in each trial provides evidence about BLUE's

Table 3
Model fits in Experiment 3. *'Intercept', 'Condition', 'Trial type', and 'Interaction' show the posterior means of each predictor along with 95% credible intervals. We contrast-coded the condition (no prior = −1, with prior = 1) and trial type (helping = −1, hindering = 1) predictors. Both models included a random intercept for each participant.*

Agent	Intercept	Condition	Trial type	Interaction
RED	58.25 [54.95, 61.48]	14.43 [11.16, 17.75]	−17.10 [−18.74, −15.44]	5.86 [4.29, 7.44]
BLUE	54.73 [52.31, 57.15]	−4.78 [−7.23, −2.25]	25.40 [23.61, 27.18]	−5.98 [−7.87, −4.12]

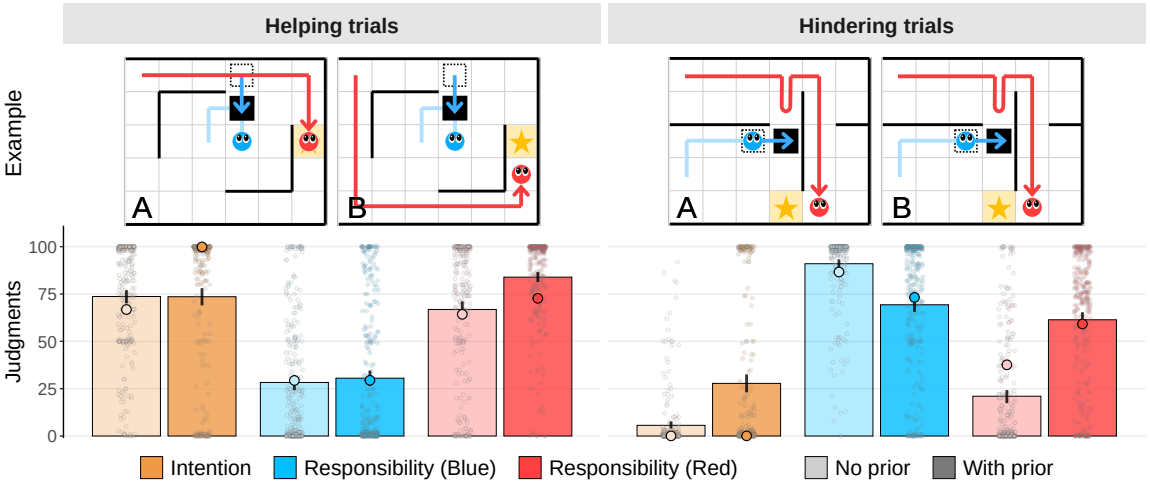

Figure 7. **Results in Experiment 3**. Participants' judgments (intention, responsibility for BLUE, and responsibility for RED) across all trials of each type in each condition. Participants in the *with prior* condition saw both rounds A and B of each trial, while those in the *no prior* condition only saw round B. All judgments were made about round B. Bars show mean ratings, error bars are bootstrapped 95% confidence intervals, large points show model predictions, and small points show individual participant judgments.

intentions and creates different normative expectations for how RED should have rationally acted in the second round. In helping trials, RED should have repeated its actions given evidence that BLUE is a helper. In hindering trials, RED should have acted differently given evidence that BLUE is a hinderer. We therefore predicted that participants in the *with prior* condition would hold RED more responsible in the second round because RED failed to act according to these expectations. Participants in the *no prior* condition, who did not see the first round, would not have formed such expectations. We additionally predicted that BLUE would generally be held less responsible in helping trials because it did not intend the failure, and that responsibility would shift to RED to account for this.

Figure 7 shows participants' intention and responsibility judgments grouped by trial type. As predicted, responsibility judgments to both agents were credibly affected by trial type, condition, and their interaction (Table 3). Participants held RED more responsible and BLUE less responsible in the *with prior* condition (RED: 72.7 on a scale of 0-100, 95% confidence interval (CI): [70.0, 75.2]; BLUE: 49.9, CI [46.6, 53.2]) compared to the *no prior* condition (RED: 44.0, CI [40.7, 47.4]; BLUE: 59.5, CI [56.2, 62.7]). RED was blamed more for failing when there was evidence that it had previously succeeded with BLUE's help or previously been hindered by BLUE in the same manner. These differences between conditions were more pronounced in hindering trials than in helping trials. Participants also generally held BLUE less responsible in helping trials (29.5, CI [26.6, 32.3]) than in hindering trials (79.7, CI [77.3, 82.1]), and conversely held RED more responsible in helping trials (75.7, CI [73.2, 78.1]) than in hindering trials (42.0, CI [39.1, 45.2]).

We modeled inferences about BLUE's intentions in the second round of each trial by either starting with uniform priors (*no prior* condition), or using posteriors from the first

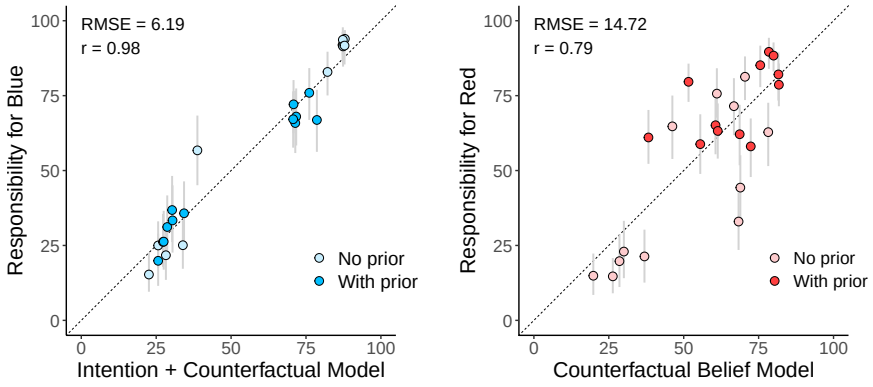

Figure 8. **Responsibility model predictions in Experiment 3**. Participants' mean responsibility judgments for BLUE (left) and RED (right) compared to model predictions. The model for BLUE combines participants' inferences about BLUE's intentions and counterfactual model predictions about what would have if BLUE hadn't been there. The model for RED predicts responsibility based on what would have happened if RED had held different prior beliefs about BLUE's intentions. Error bars are bootstrapped 95% confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

round as priors (*with prior* condition). The intention model captures participants' intention judgments well (Pearson $r = 0.88$, RMSE = 23.47). Intention judgments together with counterfactual model predictions about how likely RED would have succeeded without BLUE explain responsibility judgments for BLUE well (Figure 8), consistent with the results from Experiment 1 and 2. The posteriors on both the intention ($M = 0.61$, 95% credible interval (CrI) [0.57, 0.65]) and counterfactual ($M = 0.16$, CrI [0.12, 0.20]) predictors were positive and credible. One puzzling result is the pattern of intention judgments for hindering trials in the *with prior* condition. Surprisingly, participants were more divided in their inferences about BLUE's intentions after seeing their actions twice. Nonetheless, the intention inferences translated into responsibility judgments in the expected direction. The more likely a participant inferred that BLUE intended to help, the less responsible they held BLUE for RED's failure to reach the goal ($r = -0.77$).

We next modeled responsibility judgments for RED. RED's intentions to reach the star are unambiguous, so we focused exclusively on the counterfactual component. Specifically, we considered whether RED would have succeeded if it had held different initial beliefs about BLUE's intentions. For the *no prior* condition, we simulated a sophisticated RED agent in the second round with uniform priors. This agent assigns the same initial probability to BLUE being a hinderer or a helper. For the *with prior* condition, we simulated a sophisticated RED initialized with posterior beliefs about BLUE's intentions from the first round, capturing the beliefs we expect a rational agent *should have* formed. We fit a softmax parameter controlling how quickly RED updates its beliefs at each timestep (Bramley, Lagnado, & Speekenbrink, 2015; Edwards & Benjamin Kleinmuntz, 1968). We predicted that RED would be held more responsible for the failure in the second round to the extent that RED would have succeeded if it had properly updated its beliefs. Counterfactual model predictions were then fit to participants' responsibility judgments via a Bayesian mixed effects model.

This model also performed well at capturing responsibility (Figure 8) and the posterior on the counterfactual predictor was positive and credible ($M = 0.74$, CrI [0.66, 0.81]).

Together, the results of Experiment 3 demonstrate that responsibility judgments can be sensitive to an agent's prior beliefs about another agent's social intentions, and that counterfactual contrasts provide a flexible mechanism for capturing responsibility not only to agents who act on others (e.g. BLUE), but also to those who are acted upon (e.g. RED). We applied a similar formalism to model responsibility for BLUE and RED while implementing different relevant counterfactual scenarios. Participants' judgments about BLUE were consistent with Experiments 1 and 2, while participants held RED more responsible when they saw evidence that RED failed to anticipate BLUE's helping or hindering actions. These findings echo the folk intuition captured in expressions such as "fool me once, shame on you; fool me twice, shame on me".

General discussion

When we explain and predict others' behavior, we naturally do so by postulating mental states such as beliefs and intentions (Baker et al., 2017; Jara-Ettinger et al., 2016, 2020). We know that others are motivated not only by physical or material goals, such as acquiring an object, but also by social goals, like helping or hindering someone else. Such social behavior can be understood as rational actions that value another agent's utility as part of one's own (Netanyahu et al., 2021; Shu et al., 2020; Ullman et al., 2009). Even children might conceptualize helping and hindering in this way (Powell, 2022; Schlingloff-Nemecz et al., 2025). Here, we argue that while intention inferences are critical for understanding helping and hindering, they are not enough. Intending to help (or hinder) is different from actually helping (or hindering). Judging whether someone actually helped (or hindered) requires understanding what difference the agent's actions made by simulating what would have happened in relevant counterfactual scenarios.

We found support for the critical role that counterfactual simulations play across three experiments. Responsibility judgments in social interactions rely on both inferences about an agent's intentions towards the other agent, and assessments of how much of a difference the agent made to the outcome. We presented a computational model of responsibility that combines inverse planning to infer intentions, and counterfactual simulation to determine causal role. Our findings are consistent with prior work showing the influence of both mental state inferences and causal attributions on responsibility judgments (Alicke, 2000; Gerstenberg et al., 2018; Langenhoff et al., 2021; Malle et al., 2014; Mao & Gratch, 2012; Sosa et al., 2021). However, they go further by formally instantiating both components as computations over a shared underlying representation.

The responsibility model extends prior work in three important ways. First, we demonstrate that counterfactual reasoning, which has been shown to underlie people's causal judgments in physical domains (Beller & Gerstenberg, 2025; Gerstenberg et al., 2021; Zhou et al., 2023), plays a critical role in how people assign responsibility for social interactions. The model presented here anchors counterfactual simulations in an intuitive theory of mind that includes explicit representations of mental states, as well as a causal understanding of how those mental states drive actions (Dennett, 1989; Jara-Ettinger et al., 2016, 2020).

Second, the model captures agents interacting with varying levels of theory of mind. These range from simply optimizing another agent's physical goals (Ullman et al., 2009),

to recursively optimizing each other's nested social and physical goals (Tejwani et al., 2021, 2022), and presentational goals independent of actual social goals (Btesh et al., 2025). Sophisticated agents generate rich social behaviors such as anticipating another agent's helping or hindering, or deceptively planning to hinder an agent by appearing to help them (Oey et al., 2023; Tan, Jara-Ettinger, & Berke, 2024). People hold sophisticated agents responsible for causally influencing another agent's beliefs (Experiment 2) and acting on mistaken beliefs (Experiment 3).

Finally, the model flexibly captures responsibility judgments to different social roles using different counterfactual contrasts. Responsibility for BLUE is captured well by considering their intentions and the counterfactual scenario in which they hadn't been there. In contrast, responsibility for RED is captured by considering how RED would have acted if they had formed different beliefs about BLUE's intentions. This relates to distinctions in how responsibility is attributed to perpetrators versus victims (Cramer et al., 2013; Gönültaş, Richardson, & Mulvey, 2021; Gray & Wegner, 2009; Gray, Young, & Waytz, 2012). Victim blaming often arises from considering how the victim (rather than the perpetrator) could have acted differently (Branscombe et al., 1996; Gönültaş et al., 2021; Gray & Wegner, 2009; Niemi, Hartshorne, Gerstenberg, Stanley, & Young, 2020), especially when their experiences are perceived to be the consequences of their own actions (Dalbert, 2009; Krulewitz & Nash, 1979; Niemi & Young, 2016). More broadly, people tend to be blamed more when they could have foreseen or prevented a negative outcome from happening (Lagnado & Channon, 2008; Malle et al., 2014). We show how this attribution may arise from thinking about particular counterfactual contrasts over a person's mental states.

Limitations and extensions

A key open question raised by this work is how people select and construe the relevant counterfactual contrasts in the first place. We modeled BLUE's causal role by considering what would have happened if it hadn't been there, and RED's causal role by considering what would have happened if it had held different prior beliefs about BLUE's intentions. But many other counterfactual situations could be considered instead. People can imagine that someone could have had different traits (Brockbank, Yang, Govil, Fan, & Gerstenberg, 2024), or been replaced by someone else – another person in the same role (Wu & Gerstenberg, 2024), a "reasonable" person (Green, 1968; Shaver, 1985), or even themselves (Alicke, Dunning, & Krueger, 2005; Goldman, 2006). Because different counterfactual contrasts can yield different causal judgments (Hart & Honoré, 1959; Menzies, 2004; Schaffer, 2005), understanding how people select the most relevant one is a key avenue for future work.

Normative expectations shape how people select counterfactuals. Often, alternatives to abnormal events come to mind more easily than to normal ones (Halpern & Hitchcock, 2015; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Kahneman & Miller, 1986; Kahneman & Tversky, 1982; Kominsky & Phillips, 2019; Macrae et al., 1993; Petrocelli et al., 2011; Phillips, Luguri, & Knobe, 2015; but see Gerstenberg & Icard, 2020; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). Thus, differing expectations about how an agent *should* have acted in a situation can lead to variation in responsibility judgments (Gerstenberg et al., 2018). Our paradigm provides a rich testbed for probing the role of different counterfactual tests for how people assign responsibility (Gerstenberg, 2024).

Finally, different counterfactual contrasts also reveal different ways of making a difference. People are sensitive to the dynamics of how events unfold (Talmy, 1988; Wolff, 2007; see also Lewis, 2000; Woodward, 2011). For example, in scenario 4 in Figure 5, RED could still have succeeded without BLUE by taking a slightly different path around the box. BLUE didn't make a difference to whether RED succeeded, but did affect *how* RED succeeded. Responsibility may be sensitive to these finer-grained aspects of causation. Intentions, for instance, may reflect the stability of a cause across varying background conditions (Gerstenberg et al., 2021; Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Vasilyeva, Blanchard, & Lombrozo, 2018). An agent who deliberately planned to bring about an outcome would likely still have done so under different circumstances, whereas an agent who accidentally caused the same outcome might not have (Bratman, 1987; Halpern & Kleiman-Weiner, 2018; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Lombrozo, 2010; Martin & Cushman, 2016; Woodward, 2006). Our framework provides a starting point for implementing and testing different aspects of causation in a social domain.

Conclusion

We present a computational framework for modeling responsibility judgments that combines two cognitive processes: mental state inference and causal attribution. By embedding these mechanisms in a formal model of rational planning, we explain how people make responsibility judgments through inferring others' intentions and evaluating their causal roles by simulating what would have happened in relevant counterfactual scenarios. Our model accounts for patterns of human judgment across a range of social interactions where one agent helps or hinders another. It also opens new avenues for investigating how people construe different counterfactual contrasts and aspects of causation in social contexts. Together, these contributions move us closer to understanding the computational principles underlying people's evaluations of others' social acts.

Methods

All experiments were approved by the Stanford Institutional Review Board and preregistered on the Open Science Framework. All participants were recruited through Prolific and compensated at a rate of $12/hour for completing a study. At the beginning of each experiment, participants gave informed consent, and at the end of each experiment they optionally shared demographic information and comments about what factors influenced their responses. All experiment materials and links to preregistrations are available at `https://github.com/cicl-stanford/helping_hindering`.

Experiment 1

Participants. We recruited 200 participants (*age*: M = 37, SD = 12; *gender*: 97 female, 93 male, 9 non-binary, 1 undisclosed; *race*: 147 White, 19 Black/African American, 22 Asian, 3 American Indian/Alaska Native, 1 Native Hawaiian/Pacific Islander, 3 Multiracial, 5 undisclosed). They were assigned to the *intention*, *counterfactual*, *effort*, or *responsibility* conditions with $n = 50$ in each.

Design. We designed 30 trials by manipulating three factors: BLUE's intention, the actual outcome, and the counterfactual outcome (see Supplementary Information for details). We manipulated scenarios such that the model's posterior on BLUE's intention χ_{BLUE} was relatively certain that BLUE helped, relatively certain that BLUE hindered, or ambiguous. RED's actual outcome was either a success or a fail, and RED's outcome in the relevant counterfactual situation would have either been a success, a failure, or close (defined as succeeding with exactly zero time steps left). For example, scenario 1 in Figure 3 shows a trial in which BLUE intended to help, the actual outcome was a success, and the counterfactual outcome would have been a fail.

Procedure. Participants were guided through instructions with an example trial and then answered four comprehension questions. They had to answer all four questions correctly in order to proceed to the main task. Otherwise, they were redirected to the beginning of the instructions. During the main task, they saw 30 trials featuring a new RED and BLUE agent each time in a randomized order.

On each trial, participants clicked through a step-by-step animation and then answered a question about what happened using a continuous slider response scale. In the *intention* condition, participants were asked "What was BLUE intending to do?" with the slider ranging from "definitely hinder RED" (0) to "unsure" at the midpoint (50) to "definitely help RED" (100). In the *counterfactual* condition, participants were asked how much they agreed with the statement that "RED [would have / would still have] succeeded if BLUE hadn't been there," with the slider endpoints labeled "not at all" (0) and "very much" (100). We used "would have" if the actual outcome was a fail and "would still have" if it was a success. In the *effort* condition, participants were asked "How much effort did BLUE exert?" with the slider ranging from "very little" (0) to "very much" (100). Finally, in the *responsibility* condition, participants were asked "How responsible was BLUE for RED's [success/fail]?" with the slider ranging from "not at all" (0) to "very much" (100). The experiment took an average of 13.4 minutes (SD = 6.1) to complete.

Experiment 2

Participants. We recruited 150 participants (*age*: M = 38, SD = 12; *gender*: 78 female, 64 male, 4 non-binary, 4 undisclosed; *race*: 102 White, 20 Black/African American, 14 Asian, 1 American Indian/Alaska Native, 7 Multiracial, 6 undisclosed). They were assigned to the *responsibility*, *counterfactual*, or *intention* condition with $n = 50$ in each.

Design. We designed 24 trials by manipulating three factors: BLUE's intention, the actual outcome, and the counterfactual outcome (see Supplementary information for details). BLUE's intention was to either help ($\chi_{\text{BLUE}} > 0$, $\rho_{\text{BLUE}} > 0$), hinder ($\chi_{\text{BLUE}} < 0$, $\rho_{\text{BLUE}} < 0$), or "fake help" ($\chi_{\text{BLUE}} < 0$, $\rho_{\text{BLUE}} > 0$). The actual outcome was either a success or fail, and the counterfactual outcome was either a success, a fail, or a close. For example, scenario 1 in Figure 5 shows a trial in which BLUE had intentions to fake help, and RED failed but would have succeeded without BLUE.

Procedure. The procedure was identical to that of Experiment 1. In each trial, participants were asked to make judgments about how responsible BLUE was for the outcome, how much they agreed that RED would have succeeded if BLUE hadn't been there, or what BLUE was intending to do, depending on the condition. On average, the experiment took 13.5 minutes (SD = 5.4) to complete.

Experiment 3

Participants. We recruited 100 participants (*age*: M = 37, SD = 13; *gender*: 48 female, 50 male, 2 undisclosed; *race*: 68 White, 16 Black/African American, 9 Asian, 4 Multiracial, 3 undisclosed). They were assigned to the *with prior* or *no prior* conditions with $n = 50$ in each.

Design. We designed two types of trials: BLUE was either a helper or a hinderer. Importantly, BLUE's intentions were consistent and it always took the same actions in both rounds. In helping trials, RED succeeded in the first round, but failed in the second round after seemingly ignoring BLUE's help. For instance, in the helping example in Figure 7, BLUE pulled a box out of RED's way in both rounds but in round B RED attempted to take a different path to the goal and failed. In hindering trials, the two rounds were effectively identical. In the hindering example in Figure 7, BLUE pushed a box into RED's way in round A, forcing RED to backtrack and fail. In round B, RED failed again in the exact same manner. We designed six trials of each type for a total of 12 trials.

Procedure. The procedure was similar to Experiments 1 and 2. After correctly answering comprehension questions following the instructions, participants were shown 12 different trials in a randomized order. In the *with prior* condition, each trial consisted of two rounds of the agents interacting (see Figure 1). Participants clicked through a step-by-step animation of the first round and were then asked "Now that you have seen this trial, what do you think RED should do in Round 2?". They responded on a continuous slider from "do the same thing" (0) to "try something different" (100). Then, they clicked through the second round and were asked three new questions. The first question read, "Now that you have seen this trial, do you think BLUE is a hinderer or a helper?" with a response slider ranging from "definitely a hinderer" (0) to "unsure" at the midpoint (50) to "definitely a helper" (100). The second question read "How responsible was BLUE for the [success/fail] in this round?", with the slider ranging from "not at all" (0) to "very much" (100). The third question was the same as the second, but asked about responsibility for RED instead. Participants had access to their response from the first round and video replays of both rounds.

The *no prior* condition was similar except that each trial only featured the second round of each pair of agents, along with the same three questions about BLUE's intentions and responsibility for both agents. The experiment took an average of 17.9 minutes (SD = 9.5) to complete.

References

Aknin, L. B., Barrington-Leigh, C. P., Dunn, E. W., Helliwell, J. F., Burns, J., Biswas-Diener, R., ... Norton, M. I. (2013). Prosocial spending and well-being: Cross-cultural evidence for a psychological universal. *Journal of Personality and Social Psychology*, *104*(4), 635–652.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

Alicke, M. D., Dunning, D., & Krueger, J. I. (Eds.). (2005). *The self in social judgment*. New York: Psychology Press.

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 30).

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), 1–10.

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

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Beckers, S., Chockler, H., & Halpern, J. Y. (2024). A causal analysis of harm. *Minds and Machines*, *34*(3), 34.

Beller, A., & Gerstenberg, T. (2025). Causation, meaning, and communication. *Psychological Review*.

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), 1654–1669.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Branscombe, N. R., Owen, S., Garstka, T. A., & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome? *Journal of Applied Social Psychology*, *26*(12), 1042–1067.

Bratman, M. (1987). *Intention, plans, and practical reason*. Harvard University Press.

Brockbank, E., Yang, J., Govil, M., Fan, J. E., & Gerstenberg, T. (2024). Without his cookies, he's just a monster: a counterfactual simulation model of social explanation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Btesh, V., Lagnado, D. A., & Gerstenberg, T. (2025). Taking others for granted: balancing personal and presentational goals in action selection. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*, 1–28.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology*, *67*(1), 135–157.

California Health & Safety Code § 1799.102. (2009).

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., & Becchio, C. (2016). Decoding intentions from movement kinematics. *Scientific Reports*, *6*(1), 37036.

Chandra, K., Li, T.-M., Tenenbaum, J. B., & Ragan-Kelley, J. (2023). Acting as inverse inverse planning. In *ACM SIGGRAPH 2023 Conference Proceedings* (pp. 1–12). Association for Computing Machinery.

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

Cramer, R. J., Gorter, E. L., Cornish Rodriguez, M. D., Clark, J. W., Rice, A. K., & Nobles, M. R. (2013). Blame attribution in court: Conceptualization and measurement of perpetrator blame. *Victims & Offenders*, *8*(1), 42–55.

Cushman, F. (2024). Computational social psychology. *Annual Review of Psychology*, *75*(Volume 75, 2024), 625–652.

Dahl, A. (2015). The developing social context of infant helping in two U.S. samples. *Child development*, *86*(4), 1080–1093.

Dalbert, C. (2009). Belief in a just world. In *Handbook of individual differences in social behavior* (pp. 288–297). New York, NY, US: The Guilford Press.

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351.

Dennett, D. C. (1989). *The intentional stance*. The MIT Press.

Drayton, L. A., & Santos, L. R. (2016). Is human prosocial behavior unique? Insights and new questions from nonhuman primates. In J. D. Greene, I. Morrison, & M. E. P. Seligman (Eds.), *Positive neuroscience* (pp. 73–88). Oxford University Press.

Edwards, W., & Benjamin Kleinmuntz. (1968). Conservatism in human information processing. In *Formal representation of human judgment* (pp. 359–369). John Wiley & Sons, Inc.

Eisenberg, N., Spinrad, T. L., & Knafo-Noam, A. (2015). Prosocial development. In *Handbook of child psychology and developmental science* (pp. 1–47). John Wiley & Sons, Ltd.

Feinberg, J. (1986). Wrongful life and the counterfactual element in harming. *Social Philosophy and Policy*, *4*(1), 145–178.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gerstenberg, T. (2022). *What would have happened? Counterfactuals, hypotheticals, and causal judgments* (preprint). PsyArXiv.

Gerstenberg, T. (2024). Counterfactual simulation in causal cognition. *Trends in Cognitive Sciences*, *28*(10), 924–936.

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*(5), 936–975.
Gerstenberg, T., & Icard, T. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.
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.
Gerstenberg, T., & Tenenbaum, J. B. (2017). *Intuitive theories* (M. R. Waldmann, Ed.). 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.
Goldman, A. I. (2006). *Simulating minds: the philosophy, psychology, and neuroscience of mindreading*. New York (N.Y.): Oxford University press.
Gönültaş, S., Richardson, C. B., & Mulvey, K. L. (2021). But they weren't being careful! Role of theory of mind in moral judgments about victim and transgressor negligence. *Journal of Experimental Child Psychology*, *212*, 105234.
Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505–520.
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124.
Green, E. (1968). The reasonable man: Legal fiction or psychosocial reality? *Law & Society Review*, *2*(2), 241.
Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020, May). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*.
Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, *66*(2), 413–457.
Halpern, J. Y., & Kleiman-Weiner, M. (2018). Towards formal definitions of blameworthiness, intention, and moral responsibility. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).
Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *British Journal for the Philosophy of Science*, *56*(4), 843–887.
Hart, H. L. A., & Honoré, T. (1959). *Causation in the law*. Clarendon Press.
Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., ... Tracer, D. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795–815.
Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.
Hitchcock, C., & Knobe, J. (2009). Cause and Norm. *Journal of Philosophy*, *106*(11), 587–612.
House, B. R., Silk, J. B., Henrich, J., Barrett, H. C., Scelza, B. A., Boyette, A. H., ... Laurence, S. (2013). Ontogeny of prosocial behavior across diverse societies. *Proceedings*

of the National Academy of Sciences, *110*(36), 14586–14591.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93.

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). 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*(8), 589–604.

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

Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Kahneman, D., & Miller, D. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty* (1st ed., pp. 201–208). Cambridge University Press.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 37, p. 6).

Klocksiem, J. (2012). A defense of the counterfactual comparative account of harm. *American Philosophical Quarterly*, *49*(4), 285–300.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, *43*(11).

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Krulewitz, J. E., & Nash, J. E. (1979). Effects of rape victim resistance, assault outcome, and sex of observer on attributions about rape. *Journal of Personality*, *47*(4), 557–574.

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. (2017). Causation in legal and moral reasoning. In M. R. Waldmann (Ed.), *The Oxford handbook of causal reasoning* (Vol. 1, pp. 565–602). Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*(6), 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.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

Lipe, M. (1991). Counterfactual thinking as a framework for attribution theories. *Psychological Bulletin*, *109*, 456–471.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, 101396.

Macrae, C. N., Milne, A. B., & Griffiths, R. J. (1993). Counterfactual thinking and the perception of criminal behaviour. *British Journal of Psychology*, *84*(2), 221–226.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *The Journal of Psychology*, *25*(1), 147–186.

Mao, W., & Gratch, J. (2012). Modeling social causality and responsibility judgment in multi-agent interactions. *Journal of Artificial Intelligence Research*, *44*, 223–273.

Markman, K. D., Gavanski, I., Sherman, S. J., & McMullen, M. N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology*, *29*(1), 87–109.

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133–143.

McEllin, L., Sebanz, N., & Knoblich, G. (2018). Identifying others' informative intentions from movement kinematics. *Cognition*, *180*, 246–258.

McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*, *27*(12), 1165–1179.

Menzies, P. (2004). Difference-making in context. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 139–180). The MIT Press.

Michotte, A. (1963). *The perception of causality*. Oxford, England: Basic Books.

Netanyahu, A., Shu, T., Katz, B., Barbu, A., & Tenenbaum, J. B. (2021). PHASE: Physically-grounded abstract social events for machine social perception. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(1), 845–853.

Niemi, L., Hartshorne, J., Gerstenberg, T., Stanley, M., & Young, L. (2020). Moral values reveal the causality implicit in verb meaning. *Cognitive Science*, *44*(6), e12838.

Niemi, L., & Young, L. (2016). When and why we see victims as responsible: The impact of ideology on attitudes toward victims. *Personality and Social Psychology Bulletin*, *42*(9), 1227–1242.

Oey, L. A., Schachner, A., & Vul, E. (2023). Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, *152*(2), 346–362.

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

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology*, *56*(1), 365–392.

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology*, *100*(1), 30–46.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Powell, L. J. (2022). Adopted utility calculus: Origins of a concept of social affiliation. *Perspectives on Psychological Science*, *17*(5), 1215–1233.

Purves, D. (2019). Harming as making worse off. *Philosophical Studies*, *176*(10), 2629–2656.

Richens, J., Beard, R., & Thompson, D. H. (2022). Counterfactual harm. In *Advances in neural information processing systems* (Vol. 35, pp. 36350–36365).

Rossi, G., Dingemanse, M., Floyd, S., Baranova, J., Blythe, J., Kendrick, K. H., ... Enfield, N. J. (2023). Shared cross-cultural principles underlie human prosocial behavior at the smallest scale. *Scientific Reports*, *13*(1), 6057.

Sartorio, C. (2007). Causation and responsibility. *Philosophy Compass*, *2*(5), 749–765.

Schaffer, J. (2005). Contrastive causation. *Philosophical Review*, *114*(3), 327–358.

Schlingloff-Nemecz, L., Pomiechowska, B., Tatone, D., Revencu, B., Mészégető, D., & Csibra, G. (2025). Young children's understanding of helping as increasing another agent's utility. *Open Mind*, *9*, 169–188.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.

Schulz, L., Alon, N., Rosenschein, J. S., & Dayan, P. (2023). Emergent deception and skepticism via theory of mind. *Proceedings of the First Workshop on Theory of Mind in Communicating Agents*.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer Science & Business Media.

Shu, T., Kryven, M., Ullman, T. D., & Tenenbaum, J. B. (2020). Adventures in Flatland: Perceiving social interactions under physical dynamics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 42, p. 7).

Shu, T., Peng, Y., Zhu, S.-C., & Lu, H. (2021). A unified psychological space for human perception of physical and social events. *Cognitive Psychology*, *128*, 101398.

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.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Tan, Z. Y., Jara-Ettinger, J., & Berke, M. (2024). Reasoning about knowledge in lie production. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).

Team, R. C. (2019). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

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.

Tejwani, R., Kuo, Y.-L., Shu, T., Stankovits, B., Gutfreund, D., Tenenbaum, J. B., ... Barbu, A. (2022). Incorporating rich social interactions into MDPs. In *2022 International Conference on Robotics and Automation (ICRA)* (pp. 7395–7401). Philadelphia, PA, USA: IEEE.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2009). Help or Hinder: Bayesian Models of Social Goal Inference. In *NeurIPS* (p. 9).

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018, May). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

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.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Warneken, F., & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences*, *13*(9), 397–402.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York, NY, US: Guilford Press.

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

Wolff, P. (2007). Representing causation. *Journal of experimental psychology. General*, *136*, 82–111.

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2011). Mechanisms revisited. *Synthese*, *183*(3), 409–427.

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 - BQGT '10* (pp. 1–1). Newport Beach, California: ACM Press.

Wu, S. A., & Gerstenberg, T. (2024). If not me, then who? Responsibility and replacement. *Cognition*, *242*, 105646.

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

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, E., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology*, *28*(1), 126–136.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023). Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*, *152*(8), 2237–2269.

Supplementary information

Modeling details

Each state s in our deterministic environment is a function of the locations of the RED and BLUE agents, the locations of all boxes, whether the BLUE agent is holding a box, whether the BLUE agent has already moved the box, whose turn it is, and the number of timesteps remaining. Both agents incur a constant cost $c(a)$ for each action taken at each timestep.

Naive RED planning. The naive RED agent has level $\ell = 0$ and gets a small proximity reward r_{RED} at each timestep that is a function of its shortest navigable distance from the goal, ignoring all boxes. The closer it is to the goal, the higher this reward. RED also gets a large terminal reward r^T_{RED} when the episode ends, either because it reaches the goal or because time runs out. If RED has not reached the goal by time T, then its terminal reward also depends on its distance from the goal. This reward structure encourages RED to get as close to the goal as possible, as soon as possible, even when a box makes the goal technically impossible to reach. Together, the reward function is

$$R^0_{\text{RED}}(s, a_{\text{RED}}) = r_{\text{RED}}(s, g_{\text{RED}}) + r^T_{\text{RED}}(s, g_{\text{RED}}) - c(a_{\text{RED}}). \tag{6}$$

Note that this does not depend on BLUE's actions or goals, so the MDP M^0_{RED} reduces to a ordinary MDP. That is, the naive RED computes the Q-function

$$Q^0_{\text{RED}}(s^t, a^t_{\text{RED}}) = R^0_{\text{RED}}(s^t, a^t_{\text{RED}}) + \gamma \sum_{s^{t+1}} \mathcal{T}(s^t, a^t_{\text{RED}}, s^{t+1}) \max_{a^{t+1}_{\text{RED}}} Q^0_{\text{RED}}(s^{t+1}, a^{t+1}_{\text{RED}}) \tag{7}$$

according to the Bellman Optimality Equation. RED then forms an optimal policy by selecting the highest Q-valued action in each state, i.e. $\pi^0_{\text{RED}}(s) = \arg\max_a Q^0_{\text{RED}}(s, a)$.

Naive BLUE planning. The naive BLUE agent has level $\ell = 1$ since it reasons about a naive RED at $\ell = 0$. It gets a social reward based on an estimate of RED's reward:

$$R^1_{\text{BLUE}}(s, a_{\text{BLUE}}, \chi_{\text{BLUE}}) = \chi_{\text{BLUE}} \cdot \widetilde{R}^0_{\text{RED}}(s, a_{\text{RED}}) - c(a_{\text{BLUE}}). \tag{8}$$

The sign and magnitude of χ_{BLUE} dictate BLUE's behavior towards RED. When χ_{BLUE} is positive, BLUE attempts to maximize RED's reward by engaging in helping behavior, such as pulling aside a box to enable RED to get closer to the goal (as this increases its own reward). When χ_{BLUE} is negative, BLUE tries to minimize RED's reward. BLUE computes the Q-function:

$$Q^1_{\text{BLUE}}(s, a_{\text{BLUE}}, \chi_{\text{BLUE}}) = R^1_{\text{BLUE}}(s, a_{\text{BLUE}}, \chi_{\text{BLUE}}) + \gamma V^1_{\text{BLUE}}(s', \chi_{\text{BLUE}}) \tag{9}$$

where the value function is given by

$$V^1_{\text{BLUE}}(s', \chi_{\text{BLUE}}) = \sum_{a_{\text{RED}}} \underbrace{p(a_{\text{RED}} \mid s')}_{\text{Equation 3}} \max_{a'_{\text{BLUE}}} Q^1_{\text{BLUE}}(s'', a'_{\text{BLUE}}, \chi_{\text{BLUE}}). \tag{10}$$

This is the standard Bellman Optimality Equation adapted to account for turn-taking. BLUE's action a_{BLUE} brings the state to s', then RED's action a_{RED} brings the state to s'', from which BLUE considers its next action a'_{BLUE}. BLUE predicts RED's possible actions by solving M^0_{RED} to obtain an estimate of $\widetilde{Q}^0_{\text{RED}}$, and then assumes that RED selects from a softmax of this function (Equation 3).

Sophisticated RED planning. The sophisticated RED at level $\ell = 2$ has the same reward function as the naive RED. The difference between the agents is that the sophisticated RED additionally models the naive BLUE's policy when computing its Q-function:

$$Q^2_{\text{RED}}(s, a_{\text{RED}}) = R^2_{\text{RED}}(s, a_{\text{RED}}) + \gamma V^2_{\text{RED}}(s'). \tag{11}$$

Suppose that at the previous state s^{-1}, BLUE took action a^{-1}_{BLUE} to get to the current state s. Then, the value function is computed as an expectation over estimates of BLUE's actual social goal and BLUE's next possible action given those estimates:

$$V^2_{\text{RED}}(s') = \sum_{\widetilde{\chi}_{\text{BLUE}}} \underbrace{p(\widetilde{\chi}_{\text{BLUE}} \mid s^{-1}, a^{-1}_{\text{BLUE}})}_{\text{Equation 3}} \sum_{a_{\text{BLUE}}} \underbrace{p(a_{\text{BLUE}} \mid s', \widetilde{\chi}_{\text{BLUE}})}_{\text{Equation 2}} \max_{a'_{\text{RED}}} Q^2_{\text{RED}}(s'', a'_{\text{RED}}). \tag{12}$$

RED has two softmax inverse temperature parameters that independently control the uncertainty in these estimates.

Sophisticated BLUE planning. The sophisticated BLUE agent's reward function is similar to the naive BLUE, but additionally includes a presentational social reward based on an estimate of RED's belief about BLUE's social goal.

$$R^3_{\text{BLUE}}(s, a_{\text{BLUE}}, \chi_{\text{BLUE}}, \rho_{\text{BLUE}}) = \tag{13}$$
$$\chi_{\text{BLUE}} \cdot \widetilde{R}^2_{\text{RED}}(s, a_{\text{RED}}) + \chi^P_{\text{BLUE}} \cdot p\Big(\widetilde{\chi}^t_{\text{BLUE}} = \text{sign}(\rho_{\text{BLUE}})\Big) - c(a_{\text{BLUE}}).$$

Here, $\widetilde{p}(\widetilde{\chi}^t_{\text{BLUE}} = \rho_{\text{BLUE}})$ is the sophisticated BLUE's estimate of the sophisticated RED's belief that BLUE's (who it assumes to be naive) social goal is equal to its presentational goal. For example, if ρ_{BLUE} is positive, then BLUE attempts to appear helpful by maximizing the likelihood that RED thinks its social goal $\widetilde{\chi}_{\text{BLUE}}$ is positive, regardless of what its true social goal χ_{BLUE} is. The sophisticated BLUE's Q and V functions are similar to the naive BLUE (Equations 9 and 10), but reason over a sophisticated RED.

Parameters. Both agents can move up, down, left, or right, or stay in place. All agents incur a fixed cost of 1 on each timestep. BLUE can additionally pick up, push, pull, or put down boxes, and incurs an additional cost for picking up boxes (+1 in Experiment 1, and +0.5 in Experiments 2 and 3). RED cannot take any actions involving boxes. We set RED's proximity reward function $r_{\text{RED}} = 0.5 \cdot 0.9^{d(s, g_{\text{RED}})}$ to decay exponentially as a function of $d(s, g_{\text{RED}})$, the shortest navigable distance from RED's location in state s to the goal, ignoring all boxes. RED's terminal reward function $r^T_{\text{RED}} = 20 \cdot 0.9^{d(s, g_{\text{RED}})}$ follows a similar exponential shape.

We solved for each agent's Q and V functions using model-based planning. For all experiments, we used $\chi_{\text{BLUE}} = \pm 1$ for BLUE's actual social goal, which scales an estimate of RED's reward, and $\rho_{\text{BLUE}} = \pm 13$ for the sophisticated BLUE's presentational social goal, which scales an estimate of the sophisticated RED's belief about BLUE's actual social goal (a value between 0 and 1). All agents besides the naive RED used a softmax inverse temperature of $\beta = 3$ in estimating lower-leveled agents' policies from their estimated Q functions (Equation 3). The sophisticated RED and BLUE agents used an inverse temperature of $\beta = 0.5$ in updating intention beliefs in Experiment 2 (Equation 2).

For predicting intention inferences, the model used a softmax inverse temperature of $\beta = 1.8$ in Experiment 1, $\beta = 1.2$ in Experiment 2, and $\beta = 1.0$ in Experiment 3. These

values were fit by minimizing the squared error between model predictions and mean judgments. For predicting causal attributions, we ran 1000 counterfactual simulations for each trial in all experiments. All simulations used a probability of following the observed history $p_{\text{follow}} = 0.1$ on each timestep, and an inverse temperature of $\beta = 2.3$ in Experiment 1 and $\beta = 1.2$ in Experiments 2 and 3 to capture uncertainty in agents' policies. In Experiment 3, the sophisticated RED agent used $\beta = 0.01$ for updating beliefs about BLUE's intentions. These values were again fit by minimizing the squared error between model predictions and mean judgments.

Data analysis. When fitting intention and counterfactual judgments to predict responsibility, we re-coded values to account for the outcome. Recall that the model considers how likely the counterfactual outcome would have been different from the actual outcome. If the actual outcome was a success, then we flipped counterfactual judgments (subtracted from 100) to represent how likely RED would have *failed* had BLUE not been there. We kept the raw intention judgments because the model predicts that the more clear it was that BLUE was helping, the more responsible BLUE is for the success. If the actual outcome was a fail, then we kept the raw counterfactual judgments because they correctly express how likely participants thought the outcome would have been different. We flipped intention inferences to indicate how congruent the intention was with intending failure, i.e. the more clear it was that BLUE was hindering, the more responsible it is for the fail. All mixed effects models reported in this paper were written in Stan (Carpenter et al., 2017) and specified with the `brms` package (Bürkner, 2017) in `R` (Team, 2019).

Experiment 1 additional information

All experiments in this paper were programmed in jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023). Participants were required to answer the following comprehension questions correctly before moving on to the test trials.

1. True or False: The goal of both players is to reach the star first.
   Correct answer: False
2. Which of the following is possible in [sample scenario]? (A) RED can walk around the box. (B) RED can push the box out of the way. (C) BLUE can pull the box out of RED's way.
   Correct answer: C only
3. True or False: BLUE can either help or hinder RED using the boxes.
   Correct answer: True
4. Which of the following can pass through each other in the same grid? (A) RED and a box. (B) RED and BLUE. (C) BLUE and a box.
   Correct answer: B only

The trials covered ten different combinations of BLUE's intentions, the actual outcome, and the counterfactual outcome:

- hindering intention, actual success, counterfactual success
- hindering intention, actual fail, counterfactual success

- hindering intention, actual fail, counterfactual close
- hindering intention, actual fail, counterfactual fail
- unsure intention, actual success, counterfactual success
- unsure intention, actual fail, counterfactual fail
- helping intention, actual success, counterfactual success
- helping intention, actual success, counterfactual close
- helping intention, actual success, counterfactual fail
- helping intention, actual fail, counterfactual fail

We designed three trials for each combination for a total of 30 trials. See Figure 9 for results for all trials.

Experiment 2 additional information

The comprehension questions were identical to those in Experiment 1. The trials covered 12 different combinations of BLUE's intentions, the actual outcome, and the counterfactual outcome:

- helping intention, actual success, counterfactual success
- helping intention, actual success, counterfactual close
- helping intention, actual success, counterfactual fail
- helping intention, actual fail, counterfactual fail
- hindering intention, actual success, counterfactual success
- hindering intention, actual fail, counterfactual success
- hindering intention, actual fail, counterfactual close
- hindering intention, actual fail, counterfactual fail
- fake helping intention, actual success, counterfactual success
- fake helping intention, actual fail, counterfactual success
- fake helping intention, actual fail, counterfactual close
- fake helping intention, actual fail, counterfactual fail

A"fake helping" intention is defined as having a social goal of hindering and a presentational goal of appearing helpful. A counterfactual close outcome is a counterfactual success with exactly zero timesteps left. We designed two trials for each combination for a total of 24 trials.

See Figure 10 for results for all trials.

Experiment 3 additional information

The comprehension questions were identical to those in Experiment 1. In the *with prior* condition, we added an additional comprehension question:

5. True or False: RED and BLUE will play two rounds of the game.
   Correct answer: True

We designed 6 helping trials, where BLUE helped RED in both rounds but RED took alternative paths in the second round, and 6 hindering trials, where BLUE hindered RED in both

rounds. For participants in the *with prior* condition, judgments after the first round about what RED should do next on a scale from "do the same thing" (0) to "try something different" (100) were significantly lower in helping trials ($M = 9.6, SD = 19.8$) than in hindering trials ($M = 90.2, SD = 19.6$), $t(597.96) = -50.2$, $p < 0.001$. This suggests that participants did form strong expectations about how RED should have acted in the second round. See Figure 11 for intention and responsibility judgments for all trials.

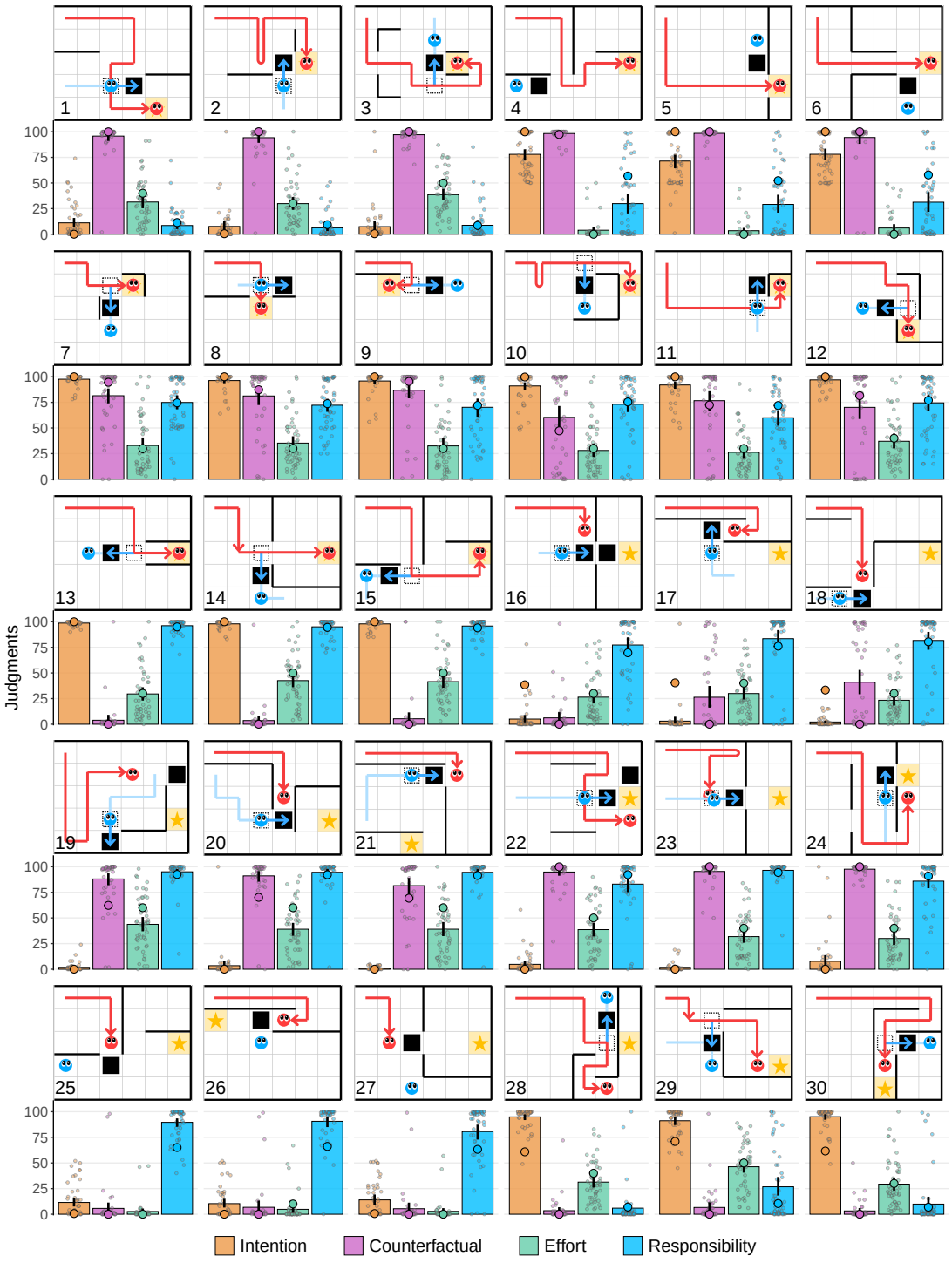

Figure 9. **Results on all trials in Experiment 1**. Participants' judgments separated by condition (counterfactual, intention, effort, and responsibility) on all trials from Experiment 1. Bars show mean ratings, error bars are bootstrapped 95% confidence intervals, large points show model predictions, and small points are individual judgments.

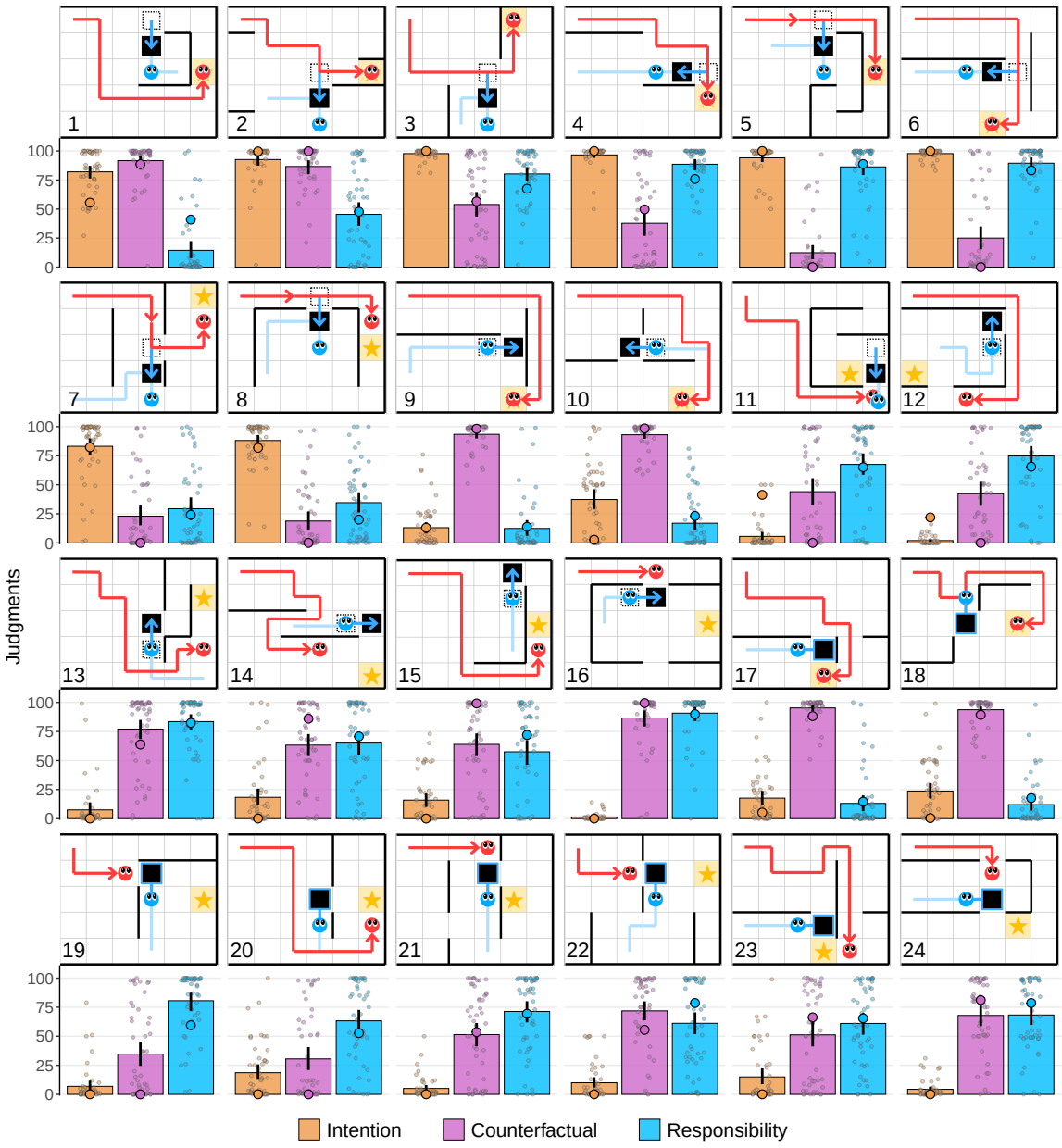

Figure 10. **Results on all trials in Experiment 2**. Participants' judgments separated by condition (counterfactual, intention, and responsibility) on all trials from Experiment 2. Bars show mean ratings, error bars are bootstrapped 95% confidence intervals, large points show model predictions, and small points are individual judgments.

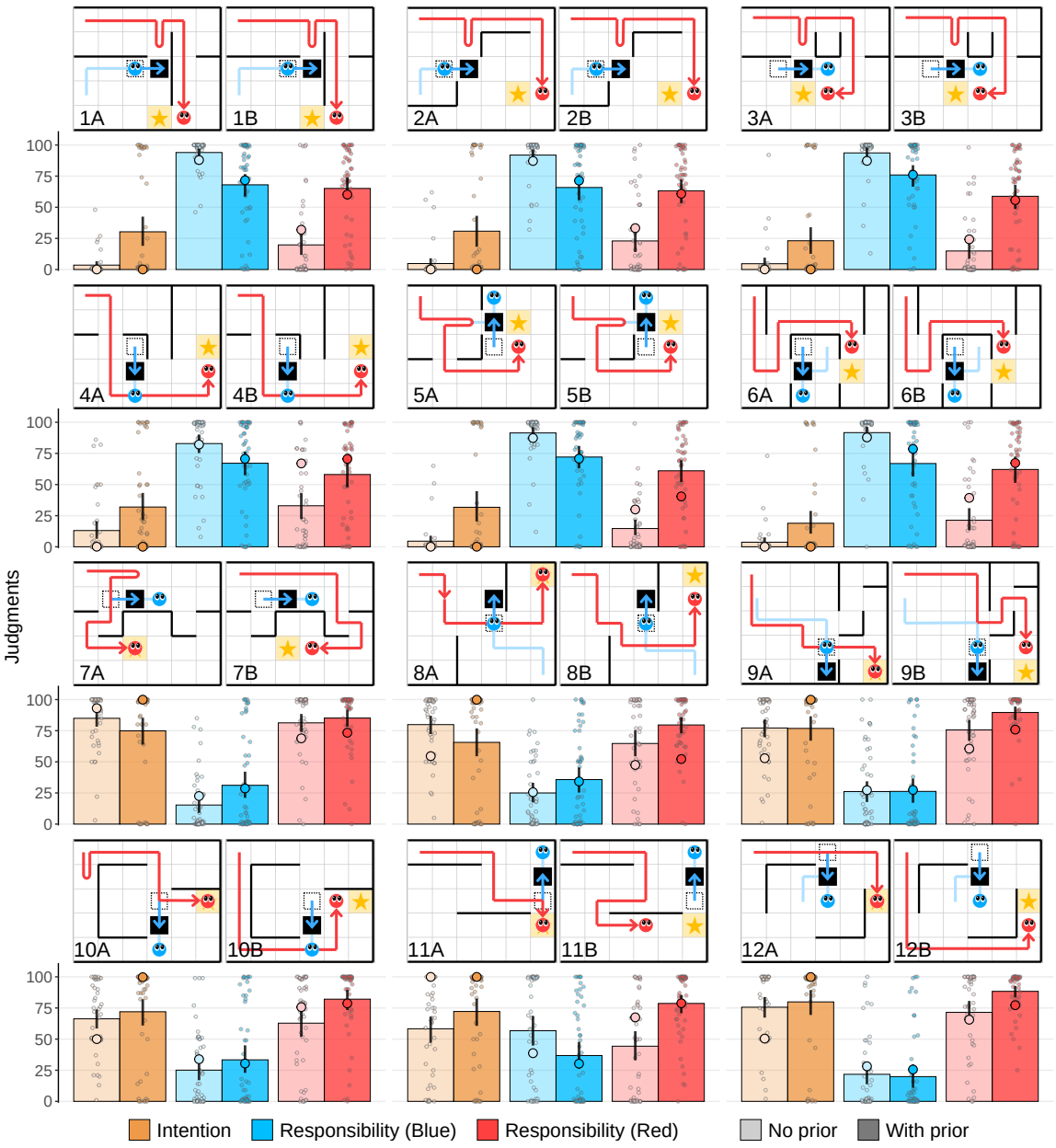

Figure 11. **Results on all trials in Experiment 3**. Participants' judgments (intention, responsibility for BLUE, and responsibility for RED) on all trials from Experiment 3, grouped by condition (*no prior* or *with prior*). Trials 1–6 are hindering trials, and 7–12 are helping trials. Bars show mean ratings, error bars are bootstrapped 95% confidence intervals, and small points show individual judgments.