A counterfactual simulation model of causal judgments about social agents

Sarah A. Wu
Department of Psychology
Stanford University
sarahawu@stanford.edu

Shruti Sridhar
Department of Computer Science
Stanford University
shrutisr@stanford.edu

Tobias Gerstenberg
Department of Psychology
Stanford University
gerstenberg@stanford.edu

Abstract—**How do people make causal judgments about others? Prior work has argued that judging causation requires going beyond what actually happened and simulating what would have happened in a relevant counterfactual situation. Here, we extend the counterfactual simulation model of causal judgments for physical events, to explain judgments about other agents' decisions and interactions. In Experiment 1, a single agent chooses what path to take to reach a goal. We find that participants' judgments about whether the agent succeeded or failed because of their decision are best explained by counterfactual judgments about whether the agent would have succeeded had they acted differently, rather than hypothetical judgments or heuristics. In Experiment 2, one agent either helps or hinders another agent from reaching the goal. Participants' judgments about whether one agent succeeded or failed because of another agent are sensitive both to what would have happened in the relevant counterfactual scenario, as well as what the other agent's actions reveal about their intentions.**

I. Introduction

How do people evaluate others' actions and decisions? From everyday occurrences like road accidents, to large-scale events like a global pandemic death toll, people attribute outcomes not only to the physical world, but also to the actions and omissions of other people [2, 14, 17, 19, 29]. Prior work has suggested that counterfactual thinking plays an important role in how people make causal judgments and explain others' actions [6, 20, 21, 22, 25, 26, 27, 32, 39]. People not only consider what someone else did, but also compare what actually happened with what would have happened had that person acted differently [12, 28]. These results suggest that causal judgments and counterfactual reasoning are intimately linked. However, little work has tried to model the cognitive processes that underlie counterfactual reasoning [but see 10] specifically as it applies to thinking about other agents.

In contrast, in the physical domain, the link between causal and counterfactual judgments has been established more firmly. Prior work has argued that people have an intuitive understanding of the physical world that is in important respects similar to the kinds of physics engines used to render realistic dynamic scenarios in computer games [9, 38]. Equipped with such a game engine in the mind, humans can make inferences about what happened in the past [11] and predictions about what will happen in the future [5, 35]. Moreover, they can use their mental model of the physical world to make causal judgments. For instance, imagine a table on which two billiard balls, ball A and ball B, collide with one another before ball B rolls through a gate. Did ball A *cause* ball B to go through the gate? Gerstenberg et al. [13] developed the counterfactual simulation model (CSM) to capture people's causal judgments in situations like these. The CSM predicts that people compare what actually happened with what they believe would have happened in relevant counterfactual scenarios. The more clear it is that ball B would have missed the gate if ball A not been there, the more people are predicted to agree that ball A caused ball B to go through the gate. The CSM yields quantitative predictions by generating noisy simulations that reflect people's uncertainty in what would have happened in the relevant counterfactual situation. These quantitative predictions are closely aligned with participants' causal judgments. Eye-tracking data further reveals that people spontaneously produce counterfactual simulations in the service of making such judgments [10].

Recently, Gerstenberg [7] investigated whether counterfactual simulations are necessary for understanding causal judgments about physical events, or whether hypothetical simulations suffice. The difference is subtle but important: a hypothetical is about a possible future (would ball B miss the gate if ball A *were not* there?), while a counterfactual is about an alternative present, and requires re-imagining past events (would ball B have missed the gate if ball A *had not been* there?). Gerstenberg found that people's causal judgments about physical events were best explained by counterfactuals rather than hypotheticals [30, 31].

Here, we build on these works by looking into situations in which people make causal judgments about psychological agents rather than physical objects. We develop a computational model of agents in a simple navigation task, and explore whether in this socially evaluative setting, causal judgments are also explained by counterfactual simulation. People often have more uncertainty about agents than objects, so it's possible that they reason about the two differently. When judging whether an object caused an outcome, people tend to imagine the counterfactual scenario in which that object had not been there. Agent behavior, on the other hand, is governed by much more than simple physical principles: it also relies on principles of rationality that dictate how mental states and abilities translate into actions given a particular situational context. Numerous counterfactual contrasts are potentially relevant – not only the scenario in which the agent had not been there, but also one in which the agent had been stronger, or smarter, or more moral, or replaced by a reasonable person instead.

Furthermore, people have a strong, automatic tendency to make social and moral evaluations when presented with agentive behavior [e.g. 16], and even 6-month-old infants show preference towards agents that are helpful towards others rather than neutral or hurtful [15]. Such social evaluation may complicate causal judgments. Causal attributions are influenced by social norms and moral considerations [1, 24]. They can also be potentially interpreted as attributions of blame, responsibility, or accountability instead, prompting consideration of other factors such as intentions that can create additional ambiguity about the causal judgment [28, 33, 36]. Here, explore how social evaluation interacts with counterfactual simulation when it comes to causal reasoning in rich multi-agent settings.

The rest of the paper is organized as follows. We first present the environments and the model, and then describe two experiments investigating causal judgments about outcomes that result from two specific counterfactual contrasts. In Experiment 1, we look at an agent's decision between two courses of action. In Experiment 2, we explore a second agent's helping or hindering interactions with the first agent.

II. Computational model

A. Environments

Our settings consist of 2D grid worlds in which agents and objects can interact. Different agents can have different goals and available actions. Here, we consider a paradigm in which one agent has 10 timesteps to reach a star, and a possible second agent can either help or hinder the first agent. On each timestep, agents can move in any of the four cardinal directions or stay in place, but cannot move through walls or black boxes. Formally, this setting can be represented as a decentralized multi-agent Markov decision process (Dec-MDP), as a tuple $\langle n, \mathcal{S}, \mathcal{A}_i, R_i, \mathcal{T} \rangle$ where n is the number of agents, $\mathcal{S}$ is the shared state space, $\mathcal{A}_i$ is the action space for agent i, $R_i : \mathcal{S} \times \mathcal{A}_i \rightarrow \mathbb{R}$ is the reward function for agent i, and $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times \ldots \times \mathcal{A}_n \rightarrow \mathcal{S}$ is the overall state transition probability.

Like the CSM, our simulation model operates over a generative model – in this case, one that dictates how agents plan and act within the bounds of the grid world. The simulation model then implements operators that allow for hypothetical and counterfactual simulations to be run.

B. Generative model

Motivated by prior work that formalizes action understanding as inverse planning in MDPs [3, 4, 18, 34, 37], we assume that humans have an intuitive psychological theory of how agents with rationality based on their mental states, their capacities, and the situational constraints. Our generative model implements this as solving the Dec-MDP, with rollouts of the resulting agent policies π_i specifying what happens over time. We record a rollout of all the policies as a history of states $\mathcal{H}^{1:T}$ where $T \leq 10$ is the length of the trial.

C. Modeling causal judgments

Given the generative model, the hypothetical and counterfactual simulation models generate predictions about what would or would have happened in alternative scenarios. We describe how they operate in turn, and then how those predictions give rise to causal judgments.

1) Hypothetical simulation

The hypothetical simulation model predicts the outcome if the agent(s) were to take different actions from the initial state. Hypothetical conditions can be thought of as alternative policies π_i'. For example, we can imagine and model how an agent would act if they were to have a different goal or different capacities (i.e. action space). The model takes the initial state and runs all π_i' forward, incorporating potential stochasticity from the transition function $\mathcal{T}$. It simulates 1000 such runs to generate a hypothetical success rate.

2) Counterfactual simulation

Counterfactual conditions can be similarly thought of as alternative policies π_i'. The counterfactual simulation model also runs all π_i' forward, but importantly conditions on any object state changes that occurred in $\mathcal{H}_{1 \ldots T}$. That is, $\mathcal{T}$ loses some stochasticity because any object state changes in the first T timesteps are no longer simulated probabilistically. Instead, all object states are maintained exactly as they are throughout $\mathcal{H}_{1 \ldots T}$. For the $(10-T)$ remaining timesteps, if any, the model resumes the original transition probabilities. The model runs 1000 simulations to generate a counterfactual success rate.

Our central research question is how people make causal judgments about outcomes that result from agents' behavior. We want to model their judgments about what happened because of e.g. the actions an agent took, or a second agent's influence. The simulation models predict that people's judgments are a function of their subjective beliefs about how likely the outcome would be or would have been *different* under relevant alternative policies. That is, they compare the actual outcome to the corresponding success rates, and make predictions based on the degree of difference.

An alternative explanation for people's causal judgments is that they don't perform any sort of mental simulation and instead consider only what actually happened. They may use properties of the observed scene as heuristics, such as how long the situation lasted (T) and what state the environment was in [40]. In our experiments, we will compare the simulation models to a heuristic model as well as one that additionally incorporates social evaluation.

D. Modeling intention judgments

In order to account for the possible influence of social evaluations, namely inferences about one agent's intention towards another based on observations of their interactions, we additionally developed an intention inference model. The intention model uses Bayesian inference to infer one agent i's intention to help or hinder agent j, given observations of agent i's actions. First, we construct the value function $\mathcal{Q}_i : \mathcal{S} \times \mathcal{A}_i \rightarrow \mathbb{R}$, which represents agent i's expected future reward for taking a given action in a given state. $\mathcal{Q}_i$ is initialized to reflect the "value" that a_i adds to agent j's pursuit of the goal. We define "value" to be a function of the change in the number of shortest paths available to agent j towards the goal, and also of the change in the length of the shortest path, such that increasing availability and decreasing length both add positive value. This represents the extent to which agent i helps agent j reach the goal by choosing a_i. There is an additional cost of 1 for any movement action and a cost of 2 for any push or pull action. Altogether, given g_i, agent i solves for its helpful or hinderful policy through Q-learning.

Next, we use the Q_i values at the end of training to infer g_i. For all $a_i \in \mathcal{A}_i$ and $s \in \mathcal{S}$, we compute

$$p(a_i \mid s, g_i) \propto \exp\big(\beta \times \mathcal{Q}_i(s, a_i)\big).$$

$$p(g_i \mid s, a_i) \;\propto\; p(a_i \mid s, g_i) \;\propto\; \exp\big(\beta \times \mathcal{Q}_i(s, a_i)\big)$$

Taking the softmax over agent i's expected reward ensures that occasional nonoptimal actions are accounted for via the free parameter β, which captures agent i's level of randomness when acting.

Given the observed state sequence $s^{1:T}$ and action sequence $a_i^{1:T}$ over a trial's T timesteps, the intention model uses Bayes' rule to compute the posterior probability of g_i:

$$p\big(g_i \mid s^{1:T}, a_i^{1:T}\big) \propto p\big(a_i^{1:T} \mid s^{1:T}, g_i\big)\; p(g_i)$$

Normalizing this posterior over all possible goals – in our settings, either help or hinder – yields the final intention prediction for agent i. In Experiment 2, we asked participants to make judgments about an agent's intentions on a continuous scale from "definitely helping" to "definitely hindering", with intermediate values reflecting more uncertainty. Thus, to quantitatively compare our inference model predictions against participants' continuous judgments, we left the $p(g_i \mid s^{1:T}, a_i^{1:T})$ as relative probabilities.

III. Experiment 1: Single agent decisions

In Experiment 1, we investigated how people may attribute outcomes to the decisions of a single agent. We tested how well causal judgments about what happened can be explained by hypothetical simulation of what would happen if the agent were to make a different decision, as well as counterfactual simulation of what would have happened had the agent made a different decision, and an alternative heuristic model.

A. Environment

In this experiment, a single agent has 10 timesteps to reach the star via one of two paths, red or blue (see examples in Figure 1). The paths may contain doors that randomly open or close with probability p_{door} on each timestep. The agent can only pass through a door if it is open, and also has a small chance p_{stall} of stalling on each timestep, which introduces some uncertainty about its behavior. Formally, we have $n = 1$, $\mathcal{A}_1 = \{\text{left, right, up, down, stay}\}$, and $\mathcal{T}$ is a function of the state $s \in \mathcal{S}$, the agent's action $a \in \mathcal{A}_1$, p_{door}, and p_{stall}.

B. Methods

1) Participants

The experiment was preregistered and posted on Prolific (hypothetical condition: https://osf.io/zw37k; counterfactual condition: https://osf.io/cxn3s; causal condition: https://osf.io/r8sdh). 150 participants (*age*: M = 35, SD = 14; *gender*: 87 female, 57 male, 2 trans male, 4 non-binary) were recruited and compensated $11/hour. They were randomly assigned to the *hypothetical*, *counterfactual*, or *causal* conditions with $n = 50$ in each.

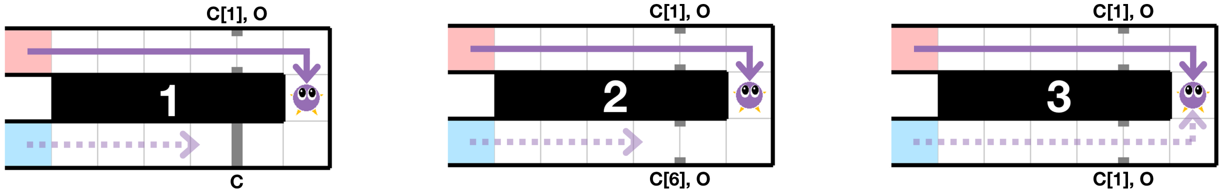
2) Procedure

Participants were introduced to the grid world setting where the agent (called the "player") took the red or blue path on each trial and then either won or lost. Both paths always looked the same initially, so there was no better or worse choice. All participants were guided through instructions with an example trial and then required to answer four comprehension questions correctly. During the main task, they saw 18 different trials in a randomized order.

In the *hypothetical* condition, participants were asked before seeing the agent's choice in each trial how much they agreed with the statement that "the player would win if they took the [color] path this time," where [color] was the color of the actual path ("red" or "blue"). Participants answered on a continuous slider from "not at all" (0) to "very much" (100) and then saw what actually happened afterward. We told them they would just be viewing feedback on their judgments in order to illustrate how often the doors randomly opened and closed. Thus, they were able to get a sense of p_{door} over the course of the experiment. The *counterfactual* condition was similar except that on each trial participants first saw everything that happened, then were asked how much they agreed that "the player would have won if they had taken the [color] path this time," where [color] was the color of the alternative path. Above the question, we displayed a video replay of what happened, which participants could re-watch as many times as they liked. The *causal* condition was identical to the counterfactual condition except the statement read, "the player [outcome] because they took the [color] path this time," where [outcome] was the actual outcome ("won" or "lost") and [color] was the color of the actual path. The experiment took an average of 10 (SD = 6) minutes to complete.

3) Design

Across the 18 trials in the experiment, we manipulated whether the agent won by reaching the star with more than one timestep left ("actual win"), just barely won or lost by exactly one timestep ("actual close"), or clearly lost by more than one timestep ("actual loss"). Similarly, we manipulated what the outcome would have been had the agent taken the alternative path ("counterfactual win", "counterfactual close",



(a) The door on the blue path stayed closed, so the agent would have lost.

(b) The door on the blue path opened after six timesteps, so the agent would have just barely succeeded.

(c) The door on the blue path opened early, and the agent would have succeeded.

Fig. 1: Diagrams of a selection of trials from Experiment 1. In all three trials, the agent took the red path and succeeded because the door on that path opened. However, what would have happened counterfactually if the agent had taken the blue path is different. The solid purple lines show the actual path and dotted lines show counterfactual paths. The doors are annotated to show state changes (e.g. `C[6],O` means the door was closed for six timesteps and then open for the rest).

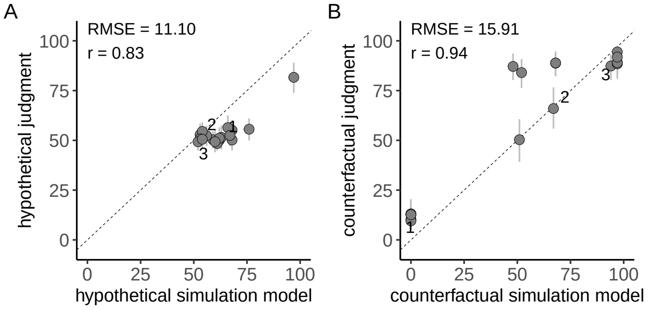


Fig. 2: Scatterplots of simulation model predictions and participants' mean judgments in the (A) hypothetical and (B) counterfactual conditions in Experiment 1. The three examples from Figure 1 are labeled. *Note*: Error bars are 95% bootstrapped confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

"counterfactual loss"). The actual path was counterbalanced.

Furthermore, we created sets of trials where what actually happened was identical, but what would have happened counterfactually, i.e. had the agent taken the alternative path, was different (see Figure 1 for an example). Thus, we would expect the same hypothetical judgments within each set about what would happen if the agent were to take the other path, but very different counterfactual judgments about what would have happened had the agent taken the other path, after the fact.

4) Modeling

For the generative model, we solve this MDP using shortest path graph search given the choice of path. The hypothetical simulation model takes the initial states of all doors and runs the generative model of the agent on the alternative path, simulating each door changing with probability p_{door} on each timestep. The counterfactual simulation model runs the generative model on the alternative path, but conditions on all door-state changes that were observed actually occurring during the T timesteps. For instance, for trial 2 in Figure 1, the model simulates the agent on the blue path with the door on that path opening after six timesteps, like it actually did. The two free parameters p_{door} and p_{stall} were fit to minimize RMSE between model predictions and participants' mean causal judgments. The optimal values were $p_{\text{stall}} = 0.12$ and $p_{\text{door}} = 0.19$.

As an alternative explanation for people's causal judgments, we also constructed a heuristic model that performs linear regression over features of the final state (at timestep T) in each trial. It considers the outcome (2 factors: win or loss) and the final states of the doors (5 factors: both open, both closed, actual open and alternative closed, actual closed and alternative open, or no doors).

C. Results & discussion

Figure 2 shows participants' mean judgments compared with corresponding simulation model predictions. The model accurately captures participants' hypothetical beliefs ($\text{RMSE}_{\text{hyp}} = 11.10$, $r_{\text{hyp}} = 0.83$), although the correlation is largely driven by the outlier. It also captures participants' counterfactual beliefs ($\text{RMSE}_{\text{cf}} = 15.91$, $r_{\text{cf}} = 0.94$), which had more range and clearly come apart from the hypotheticals. Participants were much more confident about counterfactuals they were able to see how the doors actually changed during each trial and thus no longer had uncertainty about door state changes in their simulations. Our model aligns closely with participants' judgments in both conditions, accounting for sources of uncertainty in how the environment might probabilistically change over time, and in turn how that might affect the agent's movements.

Figure 3 compares participants' mean judgments in the causal condition with predictions of the three models: hypothetical simulation, counterfactual simulation, and heuristic. For the simulation models, we directly used participants' judgments from the corresponding conditions. Causal judgments about the outcome are best explained by counterfactual judgments about what would have happened had the agent acted differently ($\text{RMSE}_{\text{cf}} = 15.67$, $r_{\text{cf}} = 0.96$). The heuristic model performs decently ($\text{RMSE}_{\text{heuristic}} = 12.34$, $r_{\text{heuristic}} = 0.9$), although it has significantly more free parameters and importantly fails to distinguish situation in which the same events happened but at critically different times, such as

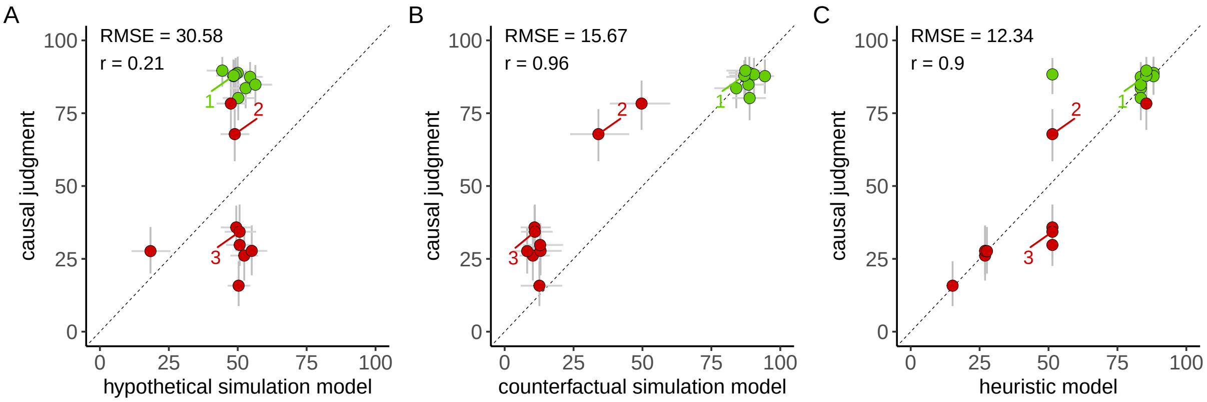


Fig. 3: Participants' mean judgments in the causal condition in Experiment 1 compared to predictions from the (A) hypothetical simulation model, (B) counterfactual simulation model, and (C) heuristic model. The green points are trials in which the counterfactual outcome would have been different from the actual outcome, and the red points are those that would have been the same. The three examples from Figure 1 are labeled. *Note*: Error bars are 95% bootstrapped confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

trials 2 and 3 (see Figure 3C). In more complex situations with multiple events and intricate timelines, we expect the counterfactual simulation and heuristic models to come apart more. Causal judgments did not align well with hypothetical judgments ($\text{RMSE}_{\text{hyp}} = 30.58$, $r_{\text{hyp}} = 0.21$).

In Experiment 2, we expand our setting to multiple agents and explore how people's intuitive understanding of others applies not only to how a single agent acts in the world, but also to how that agent may perceive and *inter*act with other agents. We explore how people can attribute outcomes to others, and how causal reasoning in these instances grows more complex as social reasoning also becomes involved.

IV. Experiment 2: Multi-agent interactions

Experiment 1 showed that causal judgments are best explained by counterfactual rather than hypothetical simulations. Thus, we focus only on counterfactual simulations here.

A. Environment

In this experiment, a red agent has 10 timesteps to reach the star. A second, blue agent has the ability to push or pull boxes around and thus can either help or hinder the red agent (see examples in Figure 4). Formally, we have $n = 2$, $\mathcal{A}_2 = \mathcal{A}_1 \cup$ {push, pull}, and $\mathcal{T}$ is a function of the state $s \in \mathcal{S}$, the red agent's action $a_1 \in \mathcal{A}_1$, and the blue agent's action $a_2 \in \mathcal{A}_2$. The blue agent's reward function depends on that of the red agent, such that $R_2 = \alpha R_1$ where α is a scaling factor representing the direction and strength of the blue agent's intentions. We use $\alpha = 0.5$ if helping, $\alpha = -0.5$ if hindering, and $\alpha = 0$ if neutral.

B. Methods

1) Participants

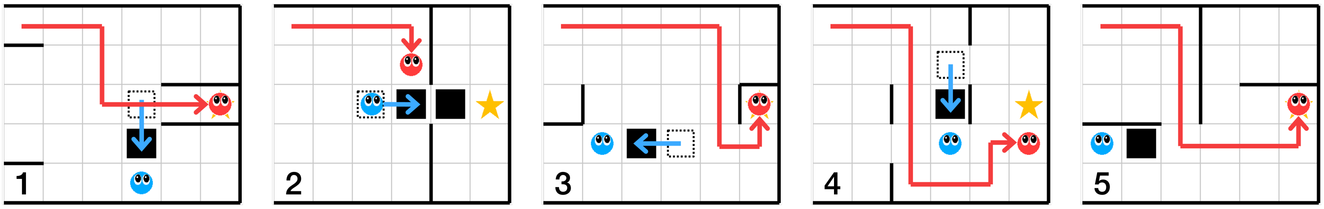
The experiment was preregistered and posted on Prolific (counterfactual condition: https://osf.io/2gekb; causal condition: https://osf.io/2w8mq; intention condition: https://osf.io/c5ah). 150 participants (*age*: M = 35, SD = 13; *gender*: 80 female, 62 male, 5 non-binary, 3 undisclosed) were recruited and compensated $11/hour. They were randomly assigned to the *counterfactual*, *causal*, or *intention* conditions with $n = 50$ in each.

2) Procedure & design

The procedure and design was similar to that of Experiment 1, featuring 24 different trials with varying combinations of actual outcomes, counterfactual outcomes, and blue agent intentions. In the *counterfactual* condition, participants saw what happened in each trial and were then asked how much they agreed that "the red player [would have / would still have] succeeded if the blue player hadn't been there." We used "would have" for trials in which the actual outcome was a fail, and "would still have" if the actual outcome was a success. Participants answered on a continuous slider from "not at all" (0) to "very much" (100). The *causal* condition was similar except that the statement read, "the red player [outcome] because of the blue player," where [outcome] was the actual outcome, either "succeeded" or "failed". Finally, in the *intention* condition, participants were asked "What was the blue player intending to do?" In this condition, they answered on a slider from "definitely hinder the red player" (0) to "definitely help the red player" (100) with the midpoint labeled "unsure" (50). The experiment took an average of 12 (SD = 8) minutes to complete.

3) Modeling

We solve for both agents' policies using Q-learning as the generative model. The counterfactual simulation runs the generative model with $\alpha = 0$ to represent the counterfactual scenario in which the blue agent had done nothing instead (as it would have zero reward). p_{stall} was again fit as a free parameter



(a) The blue agent pulled the box out of the way and the red agent succeeded. If the blue agent had done nothing, then the red agent would have failed.

(b) The blue agent hindered by pushing the box and the red agent failed, although the red agent would still have failed even if the blue agent had done nothing.

(c) The blue agent appears to potentially have intended to help by pulling the box, but made no difference to the red agent's actual path. The red agent succeeded.

(d) The blue agent appears to have attempted to help by pulling the box, but inadvertently forced the red agent to take a longer path. The red agent ultimately failed.

(e) The blue agent did nothing so their intentions here may be unclear, although they could have easily pushed the box in the red agent's way if they had intended to hinder.

Fig. 4: Diagrams of a selection of trials from Experiment 2. The red lines indicate the red agent's movements, the dashed boxes show initial box locations, and the blue arrows indicate the blue agent pushed or pulled a box to its final location.

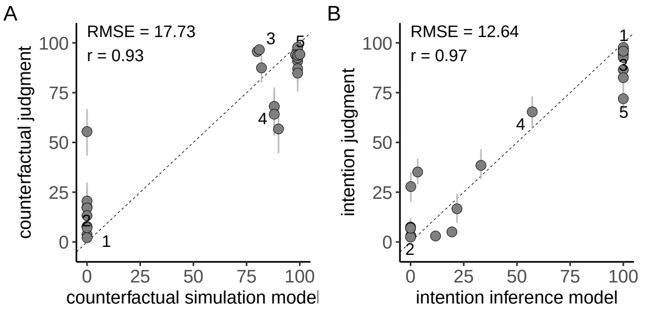


Fig. 5: Scatterplots of (A) counterfactual simulation model and (B) intention inference model predictions compared to participants' mean judgments in the corresponding conditions in Experiment 2. The examples from Figure 4 are labeled. *Note*: Error bars are 95% bootstrapped confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

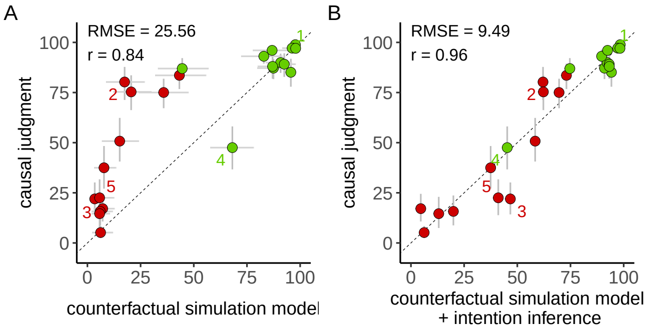


Fig. 6: Participants' mean causal judgments in Experiment 2 compared to predictions from (A) the counterfactual simulation model, and (B) a model combining both counterfactual simulation and intention inference. The green points are trials in which the counterfactual outcome would have been different from the actual outcome, and the red points are those that would have been the same. *Note*: Error bars are 95% bootstrapped confidence intervals, RMSE = root mean squared error, and r = Pearson correlation coefficient.

with an optimal value of 0.1. In the intention inference model, $\beta = 0.5$ was used.

To account for the possible influence of social evaluations, namely inferences about the blue agent's intentions, on causal judgments, we also tested a linear model that uses both counterfactual simulations and intention inferences as predictors. In the *intention* condition of the experiment, participants' responses ranged from 0 if they believed the blue agent was definitely hindering, to 100 if they believed the blue agent was definitely helping. We re-coded these values to account for the outcome. We used the raw judgments if the outcome was a success, and flipped them by subtracting from 100 if the outcome was a failure. The re-coded values thus reflect the strength or certainty of the intention in congruence to the direction of the outcome. These values, along with counterfactual model predictions, were fit as fixed effects to participants' cause judgments.

C. *Results & discussion*

Figure 5 shows participants counterfactual and intention judgments compared to corresponding model predictions. The counterfactual simulation model tends to give fairly extreme judgments (e.g. that the red agent would have succeeded without the blue agent with either 0% likelihood or over 80% likelihood), while participants were uncertain in many cases such as trial 4. The high correlation ($r_{\text{cf}} = 0.93$) is mostly driven by the two clusters of points near 0 (in which both participants and model agreed that the red agent would not have succeeded) and 100 (in which both participants and model were highly confident the red agent would have succeeded). Currently, the only gradedness in model predictions comes from p_{stall}. Future work will focus on identifying other sources of uncertainty in participants' judgments in order to improve the model and soften its predictions.

The intention inference model captures participants' judgments about the blue agent's intentions well ($r_{\text{int}} = 0.97$), including cases in which intentions were unclear. For instance, in trial 4 (see Figure 4(d)), the blue agent appears to have attempted to help by moving a box that was blocking a possible path for the red agent, but because this action was ultimately worse for the red agent, there is some ambiguity in interpretation. Both participants and model were uncertain about this trial and erred slightly on the side of helping. There is a cluster of trials in which the intention model inferred strongly that the blue agent was helping, but participants were less confident. This includes trial 5, in which the blue agent actually did nothing, but one could interpret such inaction as being relatively more consistent with a helping intention than a hindering intention. In recent work, Gerstenberg and Stephan [8] extended the counterfactual simulation model to account for people's causal judgments about *omissions* in physical events. It would be interesting to investigate similarities and differences in judgments about *inactions* in social contexts. Future work also includes more closely analyzing trials where the blue agent's intentions are not well captured by the inference model, in order to try to better understand participants' reasoning in those scenarios.

Figure 6 compares participants' mean causal judgments with predictions of the counterfactual simulation model and combined counterfactual intention model. Causal judgments are better explained by the combined model, especially trials in which the actual and counterfactual outcomes are different (red points) but participants nevertheless gave high causal ratings. This suggests that participants' judgments about whether the blue agent caused the red agent to succeed or fail are not only determined by what would have happened in a counterfactual scenario without the blue agent, but also influenced by the perceived intentions of the blue agent towards the red agent. For instance, in trial 2, the blue agent pushed a box in the red agent's way, although there was already a box preventing the red agent from reaching the goal (see Figure 4). Both participants and the counterfactual simulation model judged the red agent to be unlikely to have succeeded counterfactually, presumably due to the box already blocking the goal, as shown in Figure 5A. However, both participants and the intention inference model also perceived the blue agent as strongly hindering, as shown in Figure 5B. Given that the actual outcome was negative, these two predictors together drove the causal judgment up in Figure 6B.

V. Conclusion

How do people make causal judgments about other people's actions and interactions? In this paper, we developed a computational model that uses simulations to predict people's hypothetical and counterfactual judgments about agents' behaviors in simple grid environments. The results of Experiment 1 demonstrate that these two types of judgments come apart, that our simulation model captures the range and uncertainty in participants' responses, and that participants' causal judgments about the outcome resulting from a single agent's decision were best explained by counterfactual judgments about what would have happened had the agent acted differently. In Experiment 2, we extended our setting to multiple agents. We showed that participants' causal judgments about one agent's outcome following a second agent's helping or hindering are influenced both by counterfactual judgments about what would have happened without the second agent, and by evaluations of the second agent's intentions.

While Gerstenberg et al. [13] had shown that a counterfactual simulation model accurately captures people's causal judgments about physical events, here we build on this work by applying it to a novel domain. People not only have an intuitive understanding of how the physical world works, they also have an intuitive understanding of how other people work [4, 9, 18, 23]. Instead of considering what would have happened if an object had not been present in the scene, we show that it is also possible to simulate what would have happened if an *agent* had not been present, or had acted differently instead. This work represents initial steps towards understanding and modeling how people attribute causes to each other, and in future research we will continue to explore this in more complex and realistic ways.

Acknowledgments

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

References

[1] M. D. Alicke. Culpable causation. *Journal of Personality and Social Psychology*, 63(3):368–378, 1992.

[2] M. D. Alicke, D. R Mandel, D. Hilton, T. Gerstenberg, and D. A. Lagnado. Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science*, 10(6):790–812, 2015.

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

[4] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.

[5] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

[6] Ruth MJ Byrne. Counterfactual thought. *Annual Review of Psychology*, 67:135–157, 2016.

[7] Tobias Gerstenberg. What would have happened? Counterfactuals, hypotheticals, and causal judgments. 2022.

[8] Tobias Gerstenberg and Simon Stephan. A counterfactual

simulation model of causation by omission. *Cognition*, 216, 2021.

[9] Tobias Gerstenberg and Joshua B. Tenenbaum. Intuitive theories. In Michael Waldmannn, editor, *Oxford Handbook of Causal Reasoning*, pages 515–548. Oxford University Press, 2017.

[10] Tobias Gerstenberg, Matthew F. Peterson, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. Eye-Tracking causality. *Psychological Science*, 28(12):1731–1744, 2017.

[11] Tobias Gerstenberg, Max H. Siegel, and Joshua B. Tenenbaum. What happened? Reconstructing the past from vision and sound. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, 2018.

[12] Tobias Gerstenberg, Tomer D. Ullman, Jonas Nagel, Max Kleiman-Weiner, David A. Lagnado, and Joshua B. Tenenbaum. Lucky or clever? From expectations to responsibility judgments. *Cognition*, 177:122–141, 2018.

[13] Tobias Gerstenberg, Noah D. Goodman, David A. Lagnado, and Joshua B. Tenenbaum. A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, 128(6):936–975, 2021.

[14] York Hagmayer and Magda Osman. From colliding billiard balls to colluding desperate housewives: causal bayes nets as rational models of everyday causal reasoning. *Synthese*, 189(1):17–28, 2012.

[15] J Kiley Hamlin, Karen Wynn, and Paul Bloom. Social evaluation by preverbal infants. *Nature*, 450(7169):557–559, 2007.

[16] F. Heider and M. Simmel. An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259, 1944.

[17] Paul Henne, Laura Niemi, Ángel Pinillos, Felipe De Brigard, and Joshua Knobe. A counterfactual explanation for the action effect in causal judgment. *Cognition*, 190:157–164, 2019.

[18] Julian Jara-Ettinger, Hyowon Gweon, Laura E. Schulz, and Joshua B. Tenenbaum. The Naïve Utility Calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(10):785, 2016.

[19] Samuel G.B. Johnson and Lance J. Rips. Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, 77: 42–76, 2015.

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

[21] D. Kahneman and A. Tversky. The simulation heuristic. In D. Kahneman and A. Tversky, editors, *Judgment under uncertainty: Heuristics and biases*, pages 201–208. Cambridge University Press, New York, 1982.

[22] Lara Kirfel, Thomas F. Icard, and Tobias Gerstenberg. Inference from explanation. *Journal of Experimental Psychology: General*, 2022.

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

[24] J. Knobe and B. Fraser. Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong, editor, *Moral Psychology: The cognitive science of morality: intuition and diversity*, volume 2. The MIT Press, 2008.

[25] Jonathan F. Kominsky and Jonathan Phillips. Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, 43(11):e12792, 2019.

[26] Jonathan F Kominsky, Jonathan Phillips, Tobias Gerstenberg, David A Lagnado, and Joshua Knobe. Causal superseding. *Cognition*, 137:196–209, 2015.

[27] D. A. Lagnado, T. Gerstenberg, and R. Zultan. Causal responsibility and counterfactuals. *Cognitive Science*, 47: 1036–1073, 2013.

[28] Antonia F Langenhoff, Alexander Wiegmann, Joseph Y Halpern, Joshua B Tenenbaum, and Tobias Gerstenberg. Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, 129:101412, 2021.

[29] B. F. Malle. How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1):23–48, 1999.

[30] J. Pearl. *Causality: Models, reasoning and inference*. Cambridge University Press, Cambridge, England, 2000.

[31] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

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

[33] Jana Samland and Michael R. Waldmann. Do social norms influence causal inferences? In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, 2014.

[34] Tianmin Shu, Marta Kryven, Tomer D Ullman, and Joshua B Tenenbaum. Adventures in flatland: Perceiving social interactions under physical dynamics. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020.

[35] K. A. Smith and E. Vul. Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, 5(1):185–199, 2013.

[36] Felix A Sosa, Tomer D Ullman, Joshua B Tenenbaum, Samuel J Gershman, and Tobias Gerstenberg. Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, 217, 2021.

[37] T. D. Ullman, J. B. Tenenbaum, C. L. Baker, O. Macindoe, O. R. Evans, and N. D. Goodman. Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems*, volume 22, pages 1874–1882, 2009.

[38] Tomer D. Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B. Tenenbaum. Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences*, 21(9):649–665, 2017.

[39] G. L. Wells and I. Gavanski. Mental simulation of causality. *Journal of Personality and Social Psychology*, 56(2):161–169, 1989.

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