That was close!
A counterfactual simulation model of causal judgments about decisions

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

Abstract

How do people make causal judgments about other's decisions? 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. In our experiments, an agent chooses what path to take to reach a goal. In Experiment 1, participants either made hypothetical judgments about whether the agent would succeed were it to take a certain path, or counterfactual judgments about whether the agent would have succeeded had it taken a different path. In Experiment 2, participants made causal judgments about whether the agent succeeded or failed because of the path that it took. Our computational model accurately captured participants' judgments in both experiments and we find that causal judgments are better explained by counterfactuals rather than hypotheticals.

Keywords: causal judgment; social cognition; mental simulation; counterfactual; hypothetical.

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 (Alicke et al., 2015; Hagmayer & Osman, 2012; Malle, 1999; Henne et al., 2019; Johnson & Rips, 2015). Prior work has suggested that counterfactual thinking plays an important role in how people make causal judgments and explain others' actions (Kahneman & Tversky, 1982; Petrocelli et al., 2011; Kominsky & Phillips, 2019; Wells & Gavanski, 1989; Kirfel et al., 2022; Byrne, 2016; Kahneman & Miller, 1986; Kominsky et al., 2015; Lagnado et al., 2013). People not only consider what someone else did, but also compare what actually happened with what would have happened had that person acted differently (Gerstenberg et al., 2018; Langenhoff et al., 2021). 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 Gerstenberg et al., 2017) specifically as it applies to thinking about other agents.

The link between causal and counterfactual judgments has been established more firmly in the physical domain. 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 that are used to render realistic dynamic scenarios in computer games (Ullman et al., 2017; Gerstenberg & Tenenbaum, 2017). Equipped with such a game engine in the mind, humans can make inferences about what happened in the past (Gerstenberg, Siegel, & Tenenbaum, 2021) and make predictions about what will happen in the future (Battaglia et al., 2013; Smith & Vul, 2013). 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, Goodman, et al. (2021) 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 causal judgments (Gerstenberg et al., 2017).

Recently, Gerstenberg (2022) addressed the question of whether counterfactual simulations are necessary for understanding people's causal judgments about physical events, or whether hypothetical simulations suffice. The difference between hypotheticals and counterfactuals is subtle but important. A hypothetical asks a question about a possible future: would ball B miss the gate if ball A weren't there? A counterfactual asks a question about an alternative present: would ball B have missed the gate if ball A hadn't been there? Gerstenberg found that, in a setting in which hypotheticals and counterfactuals came apart, people's causal judgments were best explained by counterfactual simulation rather than hypothetical simulation (Pearl, 2000, 2019).

Here, we build on this work by looking into situations in which people make causal judgments about psychological agents rather than physical objects. We develop a computational model of an agent in a simple navigation task, and explore whether in this socially evaluative setting, causal judgments are also better explained by counterfactuals rather than hypotheticals. Generally, we have more uncertainty about agents than objects, and so it's possible that the way we make causal judgments about the two is different. When judging whether an object caused an outcome, people tend to imagine the counterfactual scenario in which that object had not been there, and the CSM simulates that. 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 an agent's mental states and abilities translate into their 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 with a reasonable person instead. In this paper, we focus on a specific contrast in a simple setting: an agent's decision between two courses of action.

The rest of the paper is organized as follows. We first describe the setting and the computational model. Then, we test the model and explore people's hypothetical and counterfactual judgments in Experiment 1, and causal judgments in Experiment 2. Consistent with Gerstenberg (2022), we find that participants' causal judgments are best explained by counterfactual rather than hypothetical simulations.

Computational model

We designed a grid world in which an agent can take one of two distinct paths, red or blue, to a star (see examples in Figure 1). On each timestep, the agent can move in any of the four cardinal directions or stay in place. The agent's movements are constrained by the fixed walls of the grid, and by doors that randomly open or close with a small probability p_{door} on each timestep. The agent can pass through a door only if it is open. The agent wins if they reach the star within ten timesteps. For example, in trial 2, the agent took the red path and lost because the door on that path remained closed for all $n = 10$ timesteps. Like the CSM, our simulation model operates over a generative model – in this case a model that dictates how agents plan and make choices within the bounds of the grid world – and implements operators that allow for hypothetical and counterfactual simulations to be run. We first discuss the generative model and then the two types of simulations in turn.

Generative model

Motivated by prior work that formalizes action understanding as inverse planning (Baker et al., 2017; Jara-Ettinger et al., 2016; Baker et al., 2009), we assume that humans have an intuitive psychological theory of how agents act based on their mental states, their capacities, and the situational constraints. Our generative model formalizes this in terms of the agent solving a Markov decision process (MDP) reflective of our setting. The MDP's states include all the grid squares, and the agent's actions include the four cardinal directions as well as stalling in place. Assuming that the agent knows the true state of the world on each timestep, and that they collect a positive reward upon reaching the end state (the star), they plan a sequence of actions that maximizes their expected utility under this MDP.

We implement the generative model by representing the grid as a graph, with locations as nodes and valid actions between locations as edges. We use Dijkstra's algorithm (Dijkstra, 1959) to find the shortest path from the red or blue starting location to the star. The agent executes this path but has a small chance p_{stall} of stalling on each timestep, which introduces some uncertainty about their behavior. Additionally, the doors may probabilistically open or close on each timestep. Thus, if the agent's planned action involves passing through a door, these random events could either enable the agent's movement, or force them to stall in place. Because of these two sources of uncertainty, the MDP's transition probability distribution over possible successor states for each state-action pair is stochastic.

Hypothetical simulation model

The hypothetical simulation model predicts the agent's probability of success if they were to take a particular path. It takes the initial states of all doors and runs the generative model of the agent on the alternative path, simulating each door having a probability p_{door} of changing on each timestep. It runs 1000 such simulations to generate a hypothetical success rate. For example, in trial 2, the model would simulate the agent on the blue path, with the door on that path changing with probability p_{door} each timestep.

Counterfactual simulation model

The counterfactual simulation model conditions on all the door-state changes that actually occurred during the trial's n observed timesteps. It then predicts the agent's probability of success if they had taken the alternative path, given those changes. For the $(10 - n)$ remaining timesteps (if any), the model changes doors probabilistically. It runs 1000 simulations to generate a counterfactual success rate. For instance, in trial 2, the model would simulate the agent on the blue path with the door on that path opening after the sixth timestep, like it actually did.

Modeling causal judgments

Our central research question is how people make causal judgments about the agent's decision. Specifically, we want to model their judgments about whether the agent succeeded (or failed) because of the path they took. The counterfactual simulation model predicts that people's judgments are a function of their subjective beliefs about how likely the outcome would have been *different* had the agent taken the alternative path. That is, it compares the actual outcome to the counterfactual success rate. In trial 2, in which the agent lost on the red path but would likely have won had they taken the blue path, the model would predict a high causal rating. The hypothetical simulation model similarly compares the actual outcome to the hypothetical success rate. For trial 2, the model's

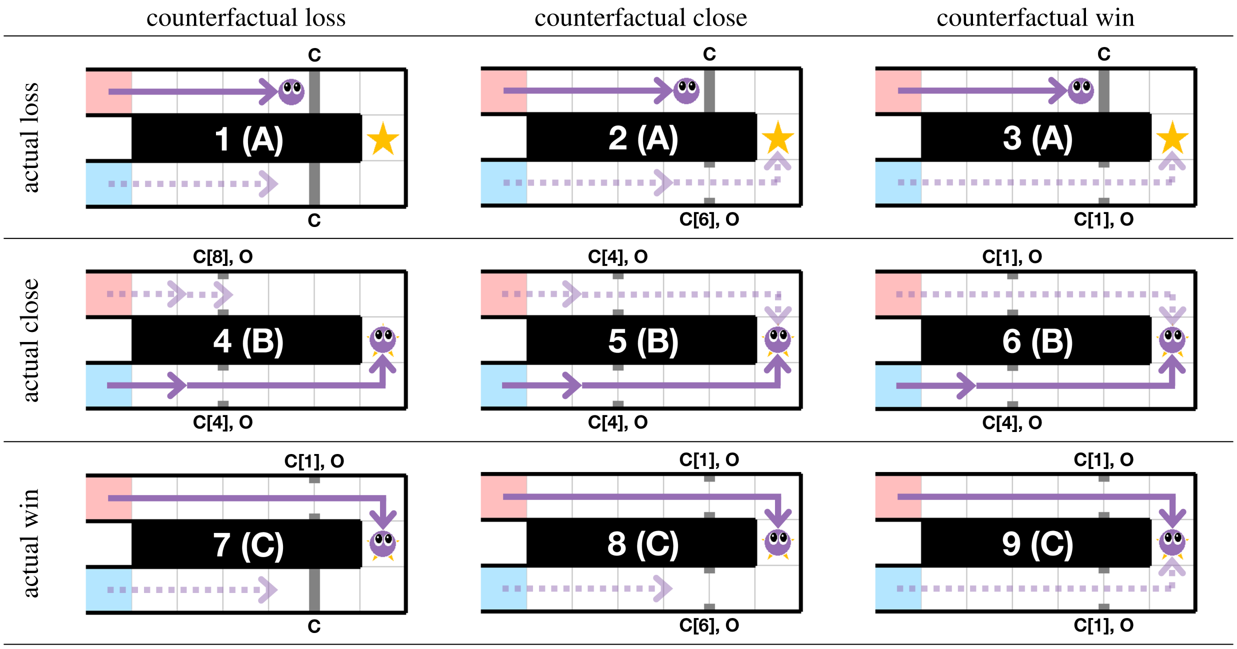

Figure 1: Diagrams of a selection of trials from both experiments. The purple arrows indicate the agent's sequence of actions, with multiple arrows meaning the agent stayed in place for at least one timestep because of a closed door. Solid lines are actual paths and dotted lines are counterfactual paths. The doors are annotated to show when they were open and closed (e.g. `C[6],O` means the door was closed for 6 timesteps and then open for the rest). Each trial is numbered, and the letters (A, B, and C) indicate triplets in which what actually happened was the same but what would have happened counterfactually is different.

prediction would depend on how often the agent might win across all hypothetical simulations.

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 (n) and what state the doors were in (White, 2014). To test this explanation, we constructed a heuristic model that performs a linear regression over visual features of the final scene. 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).

Experiment 1: Hypotheticals & counterfactuals

In Experiment 1, we investigated whether the simulation model accurately captures participants' hypothetical judgments about what would happen if the agent were to take a certain path, as well as counterfactual judgments about what would have happened had the agent taken a different path.

Methods

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

Participants The experiment was preregistered and posted as an online study on Prolific (hypothetical condition: `https://osf.io/zw37k`; counterfactual condition: `https://osf.io/cxn3s`). 100 participants (*age*: M = 33, SD = 13; *gender*: 63 female, 33 male, 1 trans male, 1 non-binary; *race*: 73 White, 10 Asian, 7 Multiracial, 5 Black, 4 Native; and 1 preferred not to say) were recruited and compensated \$11/hour. They were randomly assigned to the *hypothetical* or *counterfactual* condition with $n = 50$ in each.

Procedure Participants were introduced to the grid world setting where the agent (called the "player") had a choice of taking the red or blue path on each trial, and then either won or lost. The agent could initially see both paths, but they always looked the same, e.g. both doors were open or both were closed. Thus, the expected utility of the two paths was the same so there was no better or worse choice. Once the agent chose a path, they could not switch.

All participants were first guided through instructions with an example trial and then answered four comprehension questions to make sure they understood the setting. They were only able to proceed to the main task once they answered all four questions correctly, otherwise they were redirected to the beginning of the instructions. During the main task, they saw 18 different trials in a randomized order (see Figure 1).

In the *hypothetical* condition, at the start of each trial, participants were asked before seeing the agent's choice 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. Participants answered on a continuous slider from "not at all" (0) to "very much" (100). After answering the question, they clicked through a step-by-step play of the agent's actions and saw whether the agent ultimately won or lost. Since both paths always initially appeared the same, the choice of [color] in the hypothetical question did not matter, but we used the agent's actual choice and told participants they would just be viewing feedback on their judgments afterward. We did this in order to illustrate to participants how the doors randomly opened and closed.

The *counterfactual* condition was similar except that on each trial, participants saw everything that happened, including the agent's actions and the outcome. Then, they 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. They again answered on a slider from "not at all" (0) to "very much" (100). Displayed above the question was a looping video replay of what happened, which participants could rewatch as many times as they liked. The experiment took an average of 9.9 (SD = 7) minutes to complete.

Design Across the 18 trials in the experiment (see Figure 1 for a selection of them), 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", "counterfactual loss"). We thus had a 3×3 design with two trials for each combination. The actual path was counterbalanced, so for the two trials in each combination, the agent took the red path in one and the blue path in the other.

Furthermore, we created triplets of trials (A, B, and C) where what actually happened was identical, including the agent's actions and any door-state changes along the agent's chosen path, but what would have happened counterfactually (i.e. door-state changes along the alternative path) was different. For example, in trials 1, 2, and 3 (triplet A), the door on each path was initially closed. The agent took the red path and lost because the door on that path stayed closed for all 10 timesteps. However, the door on the blue path changed in different ways: it also stayed closed for all 10 timesteps in trial 1 (hence "counterfactual loss"), but opened just in time in trial 2 such that counterfactually the agent would have barely won ("counterfactual close"), or opened very early in trial 3 such that counterfactually the agent would have clearly won. Thus, we would expect the same hypothetical judgments across the three trials about what would happen if the agent were to take the blue path, but very different counterfactual judgments about what would have happened had the agent taken the blue path, after the fact.

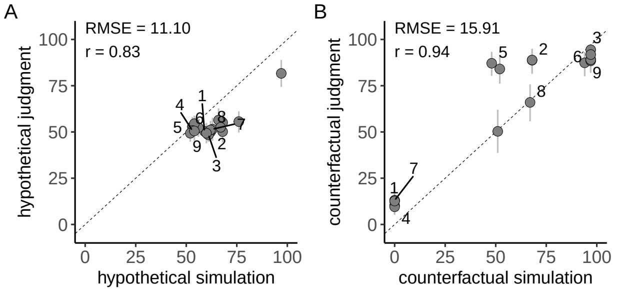

Figure 2: Scatterplot of model predictions and participants' mean judgments in the (A) *hypothetical* and (B) *counterfactual* conditions in Experiment 1. The labels on the points refer to the trials shown in Figure 1. *Note*: Error bars are 95% bootstrapped confidence intervals, RMSE = root mean squared error, *r* = Pearson correlation coefficient.

Model fitting The simulation model has two free parameters that capture sources of uncertainty in the setting. One is p_{stall}, the probability of the agent stalling on each timestep, which captures participants' uncertainty about the agent's behavior. The other is p_{door}, the probability of a door changing on each timestep, which reflects uncertainty about the environment. We fit both parameters to minimize root mean squared error between model predictions and mean judgments. The optimal values were $p_{\text{stall}} = 0.12, p_{\text{door}} = 0.19$.

Results

We discuss the results from the hypothetical and counterfactual condition in turn.

Hypothetical judgments 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 high correlation is largely driven by the outlier. In that trial, there were no doors, so participants were confident the agent would win. In the rest of the trials, both paths had one door. Participants were always unsure whether or how doors would change in each trial, so their judgments tended to cluster around the midpoint of the scale. The hypothetical simulation model accurately captures this trend.

Counterfactual judgments Participants' mean counterfactual judgments wre also accurately captured by the simulation model ($\text{RMSE}_{\text{cf}} = 15.91$, $r_{\text{cf}} = 0.94$) and had more range. For all the counterfactual loss trials, both the model and participants were confident that the agent would not have won on the other path. Similarly, for the counterfactual win trials, both model and participants assigned high likelihood that the agent would have won. Participants also thought the agent would likely have won in trials 2 and 5. In those trials, the door on the opposite path opened halfway through such that the agent would have won just in time, assuming they did not stall. However, the model gave lower ratings in those cases because it predicted that the agent would have stalled more often than participants thought it would. Finally, both the model and participants were uncertain about trial 8. In that trial, the door on the opposite path also changed just in time, but the difference is that the agent won in only 7 timesteps. Thus, there may have been additionally uncertainty about what would have happened in the remaining 3 timesteps that were never observed.

Discussion

In this experiment, we found that participants made very different hypothetical and counterfactual judgments in our setting. When there were no doors in the grid, participants were sure that the agent would win hypothetically and would have won counterfactually. However, when making hypothetical judgments on trials in which both paths had a door, participants were quite uncertain about how each door would change and what the outcome would be. In contrast, when making counterfactual judgments, participants were much more confident in almost all trials about whether the agent would have won or lost. This is because in the counterfactual condition, they were able to see how the doors actually changed during the trial, and thus no longer had uncertainty in that respect. 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.

Experiment 2: Causal judgments

In Experiment 2, we asked participants to make causal judgments about the same scenarios. We tested how well causal judgments can be explained by counterfactual simulation, hypothetical simulation, and the heuristic model.

Methods

Participants The experiment was preregistered (`https://osf.io/r8sdh`) and posted as an online study on Prolific. $n = 50$ participants (*age*: M = 40, SD = 15; *gender*: 24 female, 24 male, 1 trans male, 1 non-binary; *race*: 38 White, 6 Black, 3 Asian, 2 Multiracial, 1 preferred not to say) were recruited and compensated at a rate of $11/hour.

Procedure & Design The procedure and design were identical to that of the *counterfactual* condition in Experiment 1 except for the question being asked. In this experiment, participants were asked how much they agreed with the statement that "The player [outcome] because they took the [color] path this time." where [outcome] was either "won" or "lost" and [color] was the color of the actual path, either "red" or "blue". This was a causal judgment about the *actual* outcome unlike the previous experiment. Again, participants responded on a continuous slider from "not at all" (0) to "very much" (100). The experiment took an average of 10.9 (SD = 5) minutes to complete.

Results & Discussion

Figure 3 compares participants' mean causal judgments 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 in Experiment 1. For loss trials, we took participants' raw judgments, i.e. that the outcome hypothetically would be, or counterfactually would have been, a win instead. For trials in which the agent actually won, we reversed participants' judgments to reflect the opposite, i.e. that the agent would have lost. We now discuss the comparison with hypothetical judgments, counterfactual judgments, and the heuristic model in turn.

Causal vs. hypothetical judgments Mean hypothetical judgments are not very correlated with causal judgments ($\text{RMSE}_{\text{hyp}} = 30.58$, $r_{\text{hyp}} = 0.21$). Because participants had so much uncertainty in what would hypothetically happen in each scenario, their judgments tended to bunch near the middle of the scale, whereas causal judgments were much more varied. Hypotheticals thus cannot explain the wide range of causal judgments observed across the diverse set of trials.

Causal vs. counterfactual judgments Counterfactual judgments, on the other hand, align closely with causal judgments ($\text{RMSE}_{\text{cf}} = 15.67$, $r_{\text{cf}} = 0.96$). For most trials, judgments tend towards the extremes of the scale: participants were either quite certain that the agent won or lost because they took the baseline path, and in turn that the outcome would have been different had the agent taken the opposite path, or they were quite certain about the opposite judgment. This highlights the close relationship between causal and counterfactual reasoning – participants judged that, to the extent that the counterfactual outcome would have been different had the agent taken the opposite path, the baseline path was the cause of the outcome. This is also reflected in the coloring of the points in Figure 3B – the green trials are those in which there was a difference in outcome between the two paths, and as expected those trials generated high causal and counterfactual ratings.

Interestingly, for the red trials in which the counterfactual outcome would have been the same, mean causal judgments are consistently higher than counterfactual (contrast) judgments. This is because the counterfactual judgment is more objective in some sense, and participants strongly agreed about the counterfactual outcome in almost all trials. On the other hand, the causal question is more open to interpretation, and some participants always chose to attribute the outcome to the agent's actions in all trials, which is driving up mean causal judgments.

Heuristic A full model of the heuristic that includes outcome, door state, and their interaction as predictors failed to converge. So we fitted a separate model for wins and losses with the door state as predictor. The heuristic model does a fairly good job at predicting participants' causal judgments ($\text{RMSE}_{\text{heuristic}} = 12.34$, $r_{\text{heuristic}} = 0.9$). The model roughly captures the division between the green trials in which a counterfactual difference was made and for which we see high causal judgments as expected, and the red trials in which no difference was made. However, one limitation is that it has seven free parameters, while the two models that rely on participants judgments in the hypothetical and counterfactual condition from Experiment 1 have none. Despite its increased complexity, the heuristic model still performs worse than the

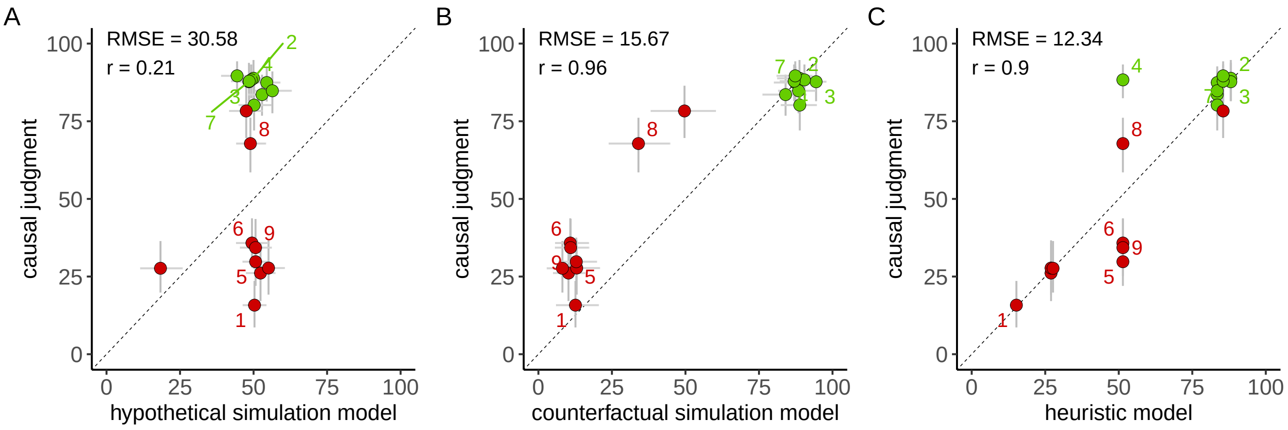

Figure 3: Participants' mean causal judgments in Experiment 2 compared to predictions from the (A) hypothetical simulation model, (B) counterfactual simulation model, and (C) heuristic model. The simulation model predictions are based on participants' judgments from Experiment 1. The green points are trials in which the counterfactual outcome would have been different from the actual outcome according to the model, and the red points are trials in which the counterfactual outcome would have been the same as what actually happened. *Note*: The labels on the points refer to the trials shown in Figure 1. Error bars are 95% bootstrapped confidence intervals. RMSE = root mean squared error, *r* = Pearson correlation coefficient.

counterfactual simulation model. One reason for this is that while the final door states indicate possible events that happened, they do not encode *when* they happened. Yet, timing is crucial for determining ultimate outcomes. For example, trials 4, 5, and 6 (triplet B) feature the same final door states, so the heuristic model predicts the same rating for all of them. However, the key difference across these trials is the exact timestep at which the door on the red path opened. Participants strongly judged the agent to have lost because they took the blue path in trial 4, where that door did not open until the last timestep, but they did not think this was the case in trials 5 or 6, where that door opened earlier. Thus, compared to the counterfactual judgments, the feature model has more free parameters and is less able to capture important details that matter for causal judgments.

General Discussion

How do people make causal judgments about other people's decisions? In this paper, we developed a computational model that uses simulations to predict people's hypothetical and counterfactual judgments about an agent's behavior in a simple grid environment. The results of Experiment 1 demonstrate that these two types of judgments come apart and that our simulation model captures the range and uncertainty in participants' responses, including uncertainty about the environment and about the agent. In Experiment 2, we found that participants' causal judgments about the outcome following the agent's actions were best explained by counterfactual judgments about what would have happened had the agent acted differently. Participants' causal judgments were also captured by a heuristic model, although this model included a number of free parameters and failed to distinguish situations in which the same events happened but at critically different times, such as the triplet of trials 4, 5, and 6 (see Figure 3C). Our setting was simple enough such that visual features of the final scene were sufficient to infer what happened, but in more complex situations with multiple events and intricate timelines, we expect the counterfactual simulation and heuristic models to come apart more strongly. Causal judgments did not align well with hypothetical judgments about what would happen if the agent were to act differently.

While Gerstenberg, Goodman, et al. (2021) 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 (Gerstenberg & Tenenbaum, 2017; Jara-Ettinger et al., 2016; Baker et al., 2017; Kleiman-Weiner et al., 2015). Instead of considering what would have happened if an object hadn't been present in the scene, here we simulate what would have happened if an agent had taken a different action. We implement people's intuitive understanding of agents in the form of a rational planning model, and assume that people can use this model to simulate counterfactuals.

Interesting theoretical questions arise in more complex settings involving multiple agents and more nuanced events. For instance, how do people reason about outcomes caused not by a single agent's actions, but by a second agent's helping or hindering (Sosa et al., 2021; Ullman et al., 2009; Shu et al., 2020)? In the future, we will explore richer settings with more agents, more graded action spaces, and more complicated event timelines that together can capture the highly uncertain nature of causal judgments about agents in real life.

Acknowledgments

This work was supported by the NSF GRFP to Sarah A. Wu, the PsychSummer program to 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

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

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

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

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.

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

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, *1*(1), 269–271.

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

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

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

Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? reconstructing the past from vision and sound. *PsyArXiv*.

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

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

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: causal bayes nets as rational models of everyday causal reasoning. *Synthese*, *189*(1), 17–28.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164.

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

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76.

Kahneman, D., & Miller, D. T. (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 & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kirfel, L., Icard, T. F., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*.

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

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), e12792.

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

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

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

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

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

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

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.

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

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

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.

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.

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

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

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