# Counterfactual simulation in causal cognition

Tobias Gerstenberg
Stanford University

May 15, 2024

Abstract

How do people make causal judgments and assign responsibility? In this paper, I show that counterfactual simulations are key. To simulate counterfactuals, we need three ingredients: a generative mental model of the world, the ability to perform counterfactual interventions on that model, and the capacity to simulate the consequences of these interventions. The counterfactual simulation model (CSM) uses these ingredients to capture people's intuitive understanding of the physical and social world. In the physical domain, the CSM predicts people's causal judgments about dynamic collision events, complex situations that involve multiple causes, omissions as causes, and causes that sustain physical stability. In the social domain, the CSM predicts responsibility judgments in helping and hindering scenarios.

*Keywords:* counterfactuals; causality; mental simulation; intuitive physics; theory of mind.

To appear in Trends in Cognitive Sciences

Tobias Gerstenberg, Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, Email: gerstenberg@stanford.edu

## From counterfactual simulations to causal judgments

> "Nothing has happened in the past; it happened in the Now. Nothing will ever happen in the future; it will happen in the Now." — Eckhart Tolle, *The Power of Now: A Guide to Spiritual Enlightenment*

Against the recommendation of many self-help books we humans spend much of our time beyond the here and now. We reminisce about the past, long for the future, and ponder how the present could have turned out differently. Usually, we cannot know what these counterfactual worlds would have looked like. But Hannes Kürmann, the protagonist in Max Frisch's stage play "Biography: A game", gets a chance to find out (32). He travels back to key decision points in his life and explores alternative paths. The result is sobering: all the counterfactual paths lead to the same miserable outcome, suggesting that our fate may be determined more by the person we are than by the choices we make.

Whether or not one agrees with the moral of Frisch's story, it's clear that counterfactual thinking plays an important role in our everyday lives. Through counterfactual thinking we make sense of how the world works, and how people work. For example, we might wonder why a driver collided with a pedestrian, or why Elizabeth passed the exam but John didn't. Part of giving an answer to why something happened is to identify what factors made a difference. Maybe the collision wouldn't have happened if the driver had been more careful? Maybe John would have passed the exam if he had studied as hard as Elizabeth? Such counterfactual thoughts are critical for answering causal questions. In this paper, I'll introduce the counterfactual simulation model (CSM) – a computational account of how counterfactual thinking underlies causal judgments and attributions of responsibility.

## Relationship status between counterfactuals and causality: "It's complicated"

While the idea of linking counterfactual thinking to causal judgments is not new (64, 65, 59, 55, 1, 123, 77, 14), it is not without its critics. Some scholars argue that counterfactuals aren't needed for capturing causation (28, 99, 121, 25), and others argue that counterfactual and causal judgments can come apart (82, 49, 79). In psychology, causal judgments are often studied by having participants read written vignettes that explicitly state what the counterfactual relationships are (e.g. 82, 69, 60, 119). But does

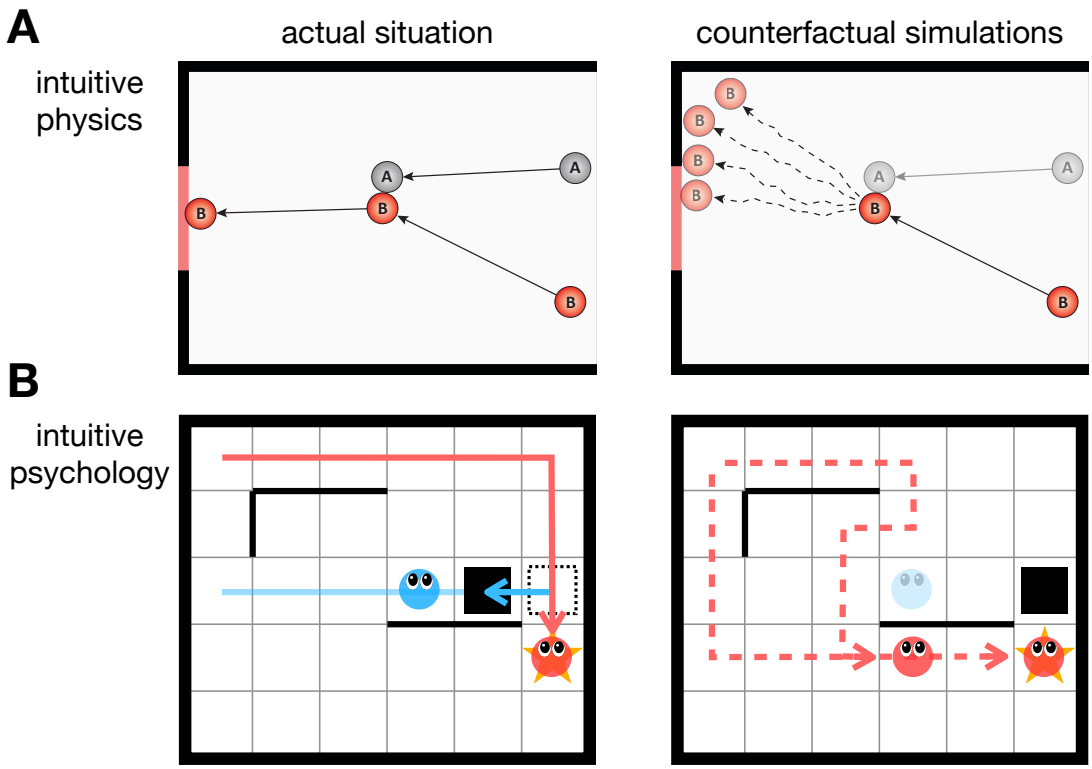


***Figure 1*. The counterfactual simulation model (CSM).** Illustration of how the CSM applies to the physical domain (**A**) and the psychological domain (**B**). The left panel shows what actually happened, and the right panel shows simulations of what could have happened in a relevant counterfactual situation. To judge whether ball A caused ball B to go through the gate, the model compares what actually happened with what would have happened if ball A hadn't been there. To judge whether the blue agent helped the red agent to reach the star in time, the model simulates what would have happened without the blue agent. Only the blue agent can move blocks, so without blue's help, red would have needed to take an alternative path and may have failed to reach the star in time. The more clear it is that the outcome would have been different in the counterfactual situation, the more causal the candidate object or agent is judged.

counterfactual information still affect people's causal judgments when they have to seek it out themselves?

I will show that causal judgments about particular events cannot be explained without relying on counterfactuals, and that dissociations between counterfactual and causal judgments arise because people care about several counterfactual contrasts when judging causation. To provide evidence that people seek out counterfactual information to answer causal questions, we use physical and social animations instead of vignettes (see also 121, 61, 107). Consider the scenarios depicted in Figure 1. We might wonder whether

ball A caused ball B to go through the gate, or whether the red agent succeeded because blue helped. Thanks to recent technological developments, such as computer game engines that represent the physical world (116, 72) and algorithms that model agent planning and decision-making (63, 3, 4), we now have the computational tools to better understand how people answer these questions (38).

The CSM predicts that people make causal judgments by using their intuitive knowledge of a domain to generate counterfactual simulations. The more certain people are that the counterfactual outcome would have been different from what actually happened, the more likely they are to say that the object or agent caused the outcome.

**Box 1: Formal frameworks for modeling causal judgment**.

Causal judgments have been captured within several formal frameworks using logical possibilities (43), statistical patterns of covariation (18, 19, 108, 109), structural causal models (50, 104, 52, 60, 92, 108, 109, 36), force vector representations (121, 122), and probabilistic programs (41, 37, 129). I'll briefly discuss two of the most prominent frameworks here.

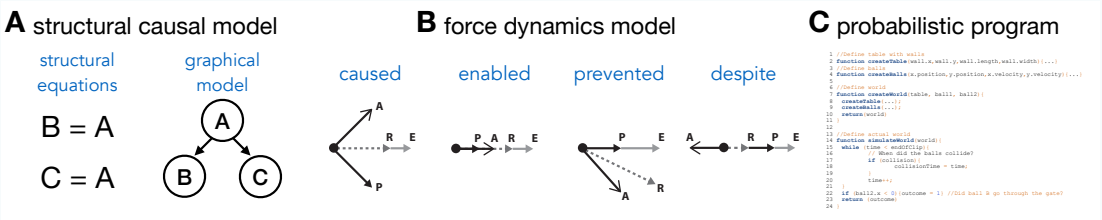


***Figure 2*. Formal frameworks for modeling causal judgments**. **A** Structural causal models express causal knowledge through a system of equations that can be visualized as graphical models. **B** The force dynamics model expresses causal knowledge as force vectors that are associated with an agent A and a patient P, and whose resultant force R may or may not lead toward endstate E. For example, "caused" is appropriate when the patient's force is not directed towards the endstate, but the agent's force combines with that of the patient to create a resultant force that leads the patient to reach the endstate. **C** Probabilistic programs express causal knowledge as a structured computer program with added randomness to capture uncertainty. The code is just for illustration.

Structural causal models represent causal knowledge as a system of structural equations that encode the relationships between variables (see Figure 2a). Structural causal models use counterfactuals to predict causal judgments. A caused B when B's value would have been different if A's value had been set to a different value. To deal with situations of causal overdetermination – where some outcome would still

have happened even if any of the individual causes hadn't happened – structural causal models consider not only whether two variables were counterfactually dependent in the actual situation, but also in other situations that could have happened (see 50, 51). While some accounts binary causal judgments (e.g., 50), others make quantitative predictions based on how close an event was to making a difference to the outcome (20, 34, 130), how abnormal or unexpected the events were (51, 52, 60), or how important different factors were a priori for the outcome to come about (74, 42, 30).

The force dynamics model (121, 122) predicts causal judgments based on the force interaction between an agent (the thing that acts) and a patient (the thing that is acted upon). Different configurations of forces map onto different causal expressions (see Figure 2b). Imagine a toy boat in a pond being pushed around by a fan. The fan "caused" the boat to hit a buoy, when the boat's force wasn't originally directed toward the buoy, but the fan exerted a force that led the boat to hit the buoy. In contrast, the fan "enabled" the boat to hit the buoy when the boat's force was already directed toward the buoy. The force dynamics model doesn't use counterfactuals. It directly maps from force configurations to causal expressions as shown in Figure 2b. The model predicts which expression best applies to a particular situation, but doesn't make quantitative predictions about how likely people would use the different expressions.

## The counterfactual simulation model

The counterfactual simulation model (CSM) combines insights from two formal frameworks for modeling causation: structural causal models and force dynamics models (see Box 1). Like in structural causal models, it uses counterfactuals to capture causation. And, like in force dynamics models, it represents the fine-grained process by which causation comes about. The CSM uses a probabilistic program to represent people's causal model of a situation (see Figure 2c, Box 1). Probabilistic programs give a detailed description of the data-generating process and include random operations to represent people's uncertainty, such that running the same stochastic process several times yields a probability distribution over possible outcomes (44, 38, 45, 33, 65, 17).

Let me illustrate how such a probabilistic program simulates counterfactuals and makes causal judgments for dynamic physical events first (5, 105, 106). Did ball A cause ball B to go through the gate in Figure 1a? The CSM predicts that people answer this question as follows (35): First, they take into account what actually happened – that

ball A collided with ball B, that ball B went though the gate, and that no other balls were present in the scene. Second, people consider a relevant counterfactual situation by imagining that ball A hadn't been present. To bring about this counterfactual situation, people "intervene" on their mental representation of the situation and remove ball A from the scene, while keeping everything else the same. Finally, they mentally simulate what the consequences of this intervention would have been. People consider where ball B would have ended up without ball A.

To run a counterfactual simulation, people use their intuitive understanding of the situation. We can model this understanding computationally as a probabilistic physics engine (116, 72; but see also 80). When people simulate where ball B would have ended up without ball A, they aren't 100% sure. To capture this uncertainty, the CSM introduces noise into the physical simulation from the timepoint onwards at which the counterfactual situation diverges from what actually happened. Here, this noise takes the form of random perturbations to ball B's velocity vector (see Figure 1a). By repeating this process, the CSM simulates how likely it is that the counterfactual outcome would have been different from what actually happened. In this example, ball B would have missed in three out of four counterfactual simulations. So, when asked whether ball A caused ball B to go through the gate, the CSM is 75% certain that it did.

The CSM not only predicts causal responsibility in physical scenarios, but also moral responsibility in social settings (126). There is a rich literature in social psychology detailing how judgments of responsibilty and blame depend on several factors, including a person's mental states (e.g., their intention and foresight) and what reasons they had for acting (e.g. 81, 100, 120, 118). What causal role a person's action played in bringing about the outcome matters, too. The CSM captures this causal role through computing counterfactual simulations. For example, to judge how responsible the blue agent was for the red agent's success in Figure 1b, the CSM simulates what would have happened if the blue agent hadn't been there. The more likely it is that the red agent (who cannot move boxes) would have failed, the more responsible the blue agent was for the red agent's success.

To sum up, the CSM predicts people's causal and responsibility judgments by computing counterfactual contrasts over a generative model (e.g., a physics engine) that captures the low-level processes (e.g., force transmission between objects) by which causes bring about effects (44, 38, 115, 22). Its predictions are probabilistic based on how certain it was that the outcome in the relevant counterfactual would have been different from

what actually happened (24). The CSM assumes that people mentally simulate what would have happened (65), and it captures people's uncertainty about the counterfactual outcome by injecting noise into its simulations. In the physical domain, that uncertainty relates to how objects would have moved, and in the psychological domain it relates to how agents would have acted.

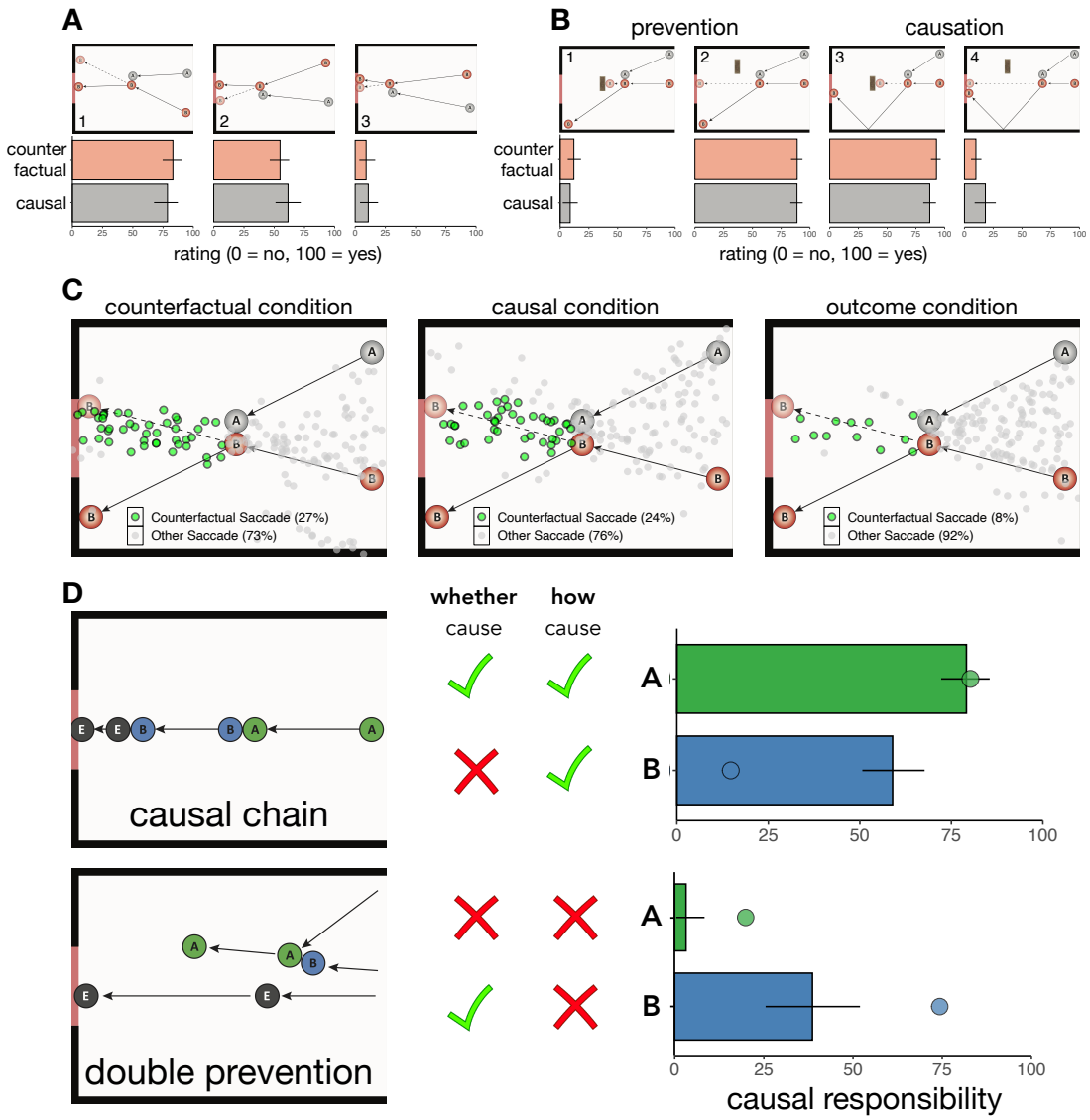


***Figure 3*. Empirical tests of the counterfactual simulation model (CSM).** **A** The CSM makes quantitative predictions about people's causal judgments (41). It predicts that people will be more likely to say that ball A caused ball B to go through the gate to the extent that they believe that ball B would have missed the gate without ball A. Here, the results show a close quantitative match between the counterfactual judgments of one group of participants (top, red), and the causal judgments of another group (bottom, gray). **B** Participants' counterfactual and causal judgments for a selection of clips from (41, Experiment 1). In clips 1 and 2, participants judged whether ball A prevented ball B from going through the gate, and in clips 3 and 4 they judged whether ball A caused ball B to go through the gate. Counterfactual and causal judgments are closely matched. Even though the causal interaction between the balls is identical in clips 1 and 2, and in clips 3 and 4, participants' causal judgments differed strongly within the pairs, suggesting that counterfactual contrasts are necessary for capturing causal judgments.

**C** Participants' eye-movements in different experimental conditions (39). Shown are saccade endpoints (fast eye-movements from one place to another) that occurred in the time between when the balls entered the scene and collided with one another. We classified as counterfactual saccades, those that ended close to the path that ball B would have taken in ball A's absence. Participants produced many more counterfactual saccades in the counterfactual and causal condition than in the outcome condition. **D** Causal responsibility judgments for three-ball scenarios (41). In the causal chain scenario, ball B is judged to be somewhat responsible for the outcome even though participants judge that ball E would have gone through the gate without ball B. In the double prevention scenario, ball B doesn't get much responsibility for ball E's going through the gate even though participants judge that ball E would have missed the gate without ball B. Together, this shows that people's causal judgments are not only sensitive to whether a candidate cause made a difference to whether the outcome happened, but also to how it came about. Note: Bars show mean judgments with 95% bootstrapped confidence intervals. Points in **D** show mean counterfactual judgments of whether ball E would have missed the gate if ball A or B hadn't been present in the scene.

### *Quantitative prediction of causal judgments*

The CSM is the first quantitative account of how people make causal judgments about dynamic collision events. Consider the physical animations shown in Figure 3a. Did ball A cause ball B to go through the gate in each of the clips? In all three clips, ball A and ball B collide with one another and ball B goes through the gate. However, what would have happened without ball A differs. In clip 1, ball B would have missed the gate, in clip 2 it's unclear, and in clip 3 ball B would likely have gone through the gate. We asked one group of participants a causal question about what happened, "Did ball A cause ball B to go through the gate?", and another group a counterfactual question, "Would ball B have missed the gate if ball A hadn't been present in the scene?" (41). As the CSM predicts, participants' answers to these two questions aligned very closely across 18 animations. Participants agreed with the causal statement to the extent that they believed that the counterfactual was true ($r = .96$).

### *Do we need counterfactuals?*

What happened in each clip in Figure 3a was slightly different. So, it's possible, in principle, that participants' judgments could be explained by a model like the force dynamics model that only takes into account what actually happened without relying on counterfactual simulations (see Box 1). To demonstrate that counterfactuals are necessary, we created situations in which what actually happened was identical, but the counterfactual outcomes differed (see 41, Experiment 1). In clips 1 and 2 in Figure 3b,

ball A and ball B interact in the same way. What differs is how ball B would have moved if ball A hadn't been present in the scene. In clip 1, ball B would have been blocked. In clip 2, it would have gone into the gate. As predicted by the CSM, participants judged that ball A prevented ball B from going through the gate in clip 2 but not in clip 1. Since the force vectors are the same in both clips, the force dynamics model cannot capture that ball A prevented ball B in one clip but not the other. Clips 3 and 4 show that the counterfactual contrast also matters when ball B ended up going through the gate. Ball A caused ball B to go through the gate in clip 3 but not in clip 4.

### *Do people spontaneously simulate counterfactuals?*

People's counterfactual and causal judgments are closely aligned across a range of situations. Can we go beyond these correlational findings and get even more direct evidence that people simulate counterfactuals when making causal judgments? Remember that the CSM computes counterfactual probabilities by running multiple simulations of what would have happened (see Figure 1a). Do people do the same? To answer this question, we tracked participants eye-movements while they were watching the video clips and asked them different questions about what happened (39). In the counterfactual condition, participants judged whether the outcome would have been different if ball A hadn't been present in the scene. In the causal condition, participants judged whether ball A caused ball B to go through the gate, or whether it prevented ball B from going through. In the outcome condition, participants judged whether ball A went right through the middle of the gate (in case it went in), or whether it completely missed the gate (in case it had missed).

Figure 3c shows the endpoints of participants' saccades (fast eye-movements from one place to another) that happened between the time when both balls entered the scene and when they collided. We classified saccades as "counterfactual saccades" (shown in green) whose endpoints were close to the path that ball B would have taken if ball A hadn't been present. Participants' eye-movements in the counterfactual and causal conditions were remarkably similar. Participants looked not only at the balls, but also where ball B would have gone had ball A not been present in the scene. In the outcome condition, participants were much less likely to produce such counterfactual saccades. So when people are asked a causal question about what happened, they spontaneously simulate counterfactual outcomes.

***What happens when there are multiple candidate causes?***

So far, we have explained people's causal judgments by assuming that they imagine what would have happened if the candidate cause hadn't been there. However, as it turns out, this is not the only counterfactual contrast that matters. People care not only about *whether* a cause made a difference but also *how* it did so (41). In the CSM, whether-causation is computed by removing the candidate cause from the scene and checking whether the outcome would have been different. How-causation is computed by applying a small change to the candidate cause (e.g., slightly perturbing its initial position) and checking whether this would have made a difference to the outcome event, finely construed.

To appreciate that both counterfactual contrasts matter, consider the causal chain scenario shown in Figure 3d. Ball A enters the scene from the right and knocks into ball B which subsequently knocks ball E into the gate. Ball B and ball E weren't moving at the beginning of the clip. When asked about how responsible ball A and ball B were for ball E's going through the gate, participants gave high ratings for ball A and relatively high ratings for ball B (see 41, Experiment 2). This illustrates that people's causal judgments are sensitive to different ways in which a candidate cause makes a difference to the outcome. Ratings for ball A were high because it was both a whether-cause and a how-cause, and lower for ball B because it was only a how-cause (ball E would have gone through the gate even if ball B had been removed, but perturbing ball B would make ball E go through the gate slightly differently).

As another example, consider the double prevention scenario in Figure 3d. Ball E enters from the right and goes through the gate without any interference. However, in the background, ball B knocks ball A out of the way. Ball A would have otherwise prevented ball E from going through the gate. This is an instance of double prevention because ball B prevents ball A from preventing ball E from going through the gate (cf., 49, 56). Here, ball B was a whether-cause but not a how-cause of the outcome, so participants gave a relatively low causal responsibility rating. The CSM's graded predictions in these scenarios result from two factors: First, it predicts higher causal ratings the more certain it is that a candidate was a whether-cause of the outcome (just like in Figure 3a and b). Second, it predicts the overall causal judgment as a weighted additive function of the different causal aspects. We found that some participants' causal judgments were more strongly affected by whether-causation, and others' by how-causation.

By considering multiple aspects of causation, the CSM also has a natural way of

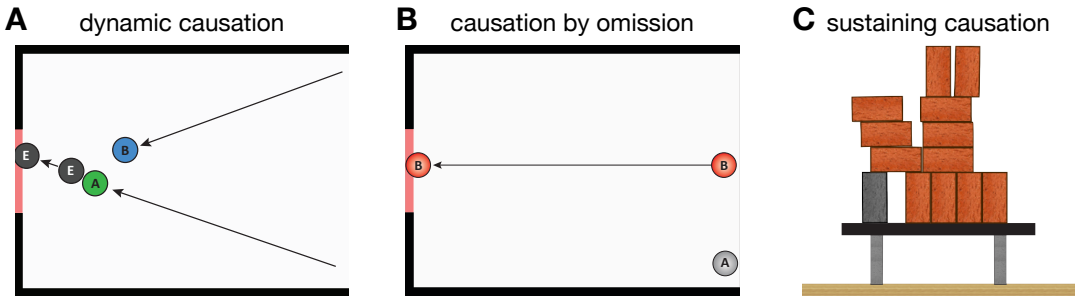


***Figure 3*. The counterfactual simulation model (CSM) captures different types of causal events**. **A** It makes predictions about people's causal judgments for dynamic collision events ("Did ball A cause ball E to go through the gate?"; 41, 39). **B** It also predicts people's judgments for omissive causes ("Did ball B go through the gate because ball A didn't hit it?"; 37). **C** And it captures people's judgments for sustaining causes ("To what extent is the black block responsible that the others stay on the table?"; 129).

differentiating between causal expressions such as "caused", "enabled", and "affected" (87, 104, 98, 104, 43, 19, 15). Instead of using force-vectors to define what each expression means (see Figure 2b, Box 1; 121), the CSM uses logical combinations of counterfactual contrasts (10, 11). Accordingly, "affected" means that a candidate was either a whether-cause or a how-cause (or both), "enabled" means that it was a whether-cause, and "caused" means that it was both a whether-cause and a how-cause. By combining this new semantics of causal expressions with a model of pragmatic inference (26, 31), the CSM accurately captured which causal expressions participants selected as the best description of what happened, and what inferences they made about what happened based on a given causal expression (10).

## Different types of causation

The CSM elucidates the cognitive mechanisms that underlie people's judgments about different types of causal relationships (see Figure 3). We have focused so far on dynamic causation events like the one in Figure 3a. Here, ball A knocks ball E into the gate before ball B would have done the same. Such situations of preemption have been used to argue against counterfactual theories of causation (119, 121). A simple counterfactual test of what would have happened if either ball hadn't been there, doesn't distinguish between the two. However, it's intuitively clear that ball A did the causing here, while ball B did nothing. The CSM partly accounts for this intuition because only ball A was a how-cause but not ball B. However, it's a problem for the CSM that participants' judgments for

ball A are at ceiling even though it wasn't a whether-cause (and generally, participants care about both aspects of causation). Better understanding people's judgments in situations of causal overdetermination remains a challenge for future work (see also Box 1).

Sometimes people cite an event that didn't happen as the cause of an outcome. Such causation by omission has attracted a lot of attention in philosophy and psychology (57, 84, 9, 58, 66, 122, 78). Did ball B go through the gate because ball A didn't hit it in Figure 3b? According to the CSM, people answer this question by simulating what would have happened if ball A had hit ball B. In this case, it's likely that ball A would have prevented ball B from going through the gate, so the answer is 'yes' (37).

People mostly talk about causation when something happens. However, causation is also at play when nothing happens. Consider the tower of blocks shown in Figure 3c. How responsible is the black block for the red blocks staying on the table? The CSM assumes that people judge responsibility by considering whether the black block prevents the others from falling, and mentally simulating what would have happened if it had been removed. Across a series of experiments, we found that the CSM accurately captured participants' judgments (129).

## Causation in the social world

Causation happens not only between billiard balls but between people, too (48, 38, 1). When people assign responsibility, they care about what causal role a person's action played in bringing about the outcome (34, 130, 74), and what the action revealed about the person (40, 75, 125, 107, 126, 68, 23).

Prior work developed computational models of what an action reveals about a person's mental states. The Bayesian Theory of Mind framework construes action understanding as inverse planning (3, 4, 63, 2, 62). People generally plan to take actions that maximize expected utility subject to their beliefs and desires (27). Because of this principled way in which mental states cause actions, an observer can reason backward from an action to the likely mental states that caused it. And because one person's utility can include another person's utility, this framework also supports inferences about social interactions, such as whether one person intended to help or hinder another (47, 91, 114, 102, 126, 21).

Intending to help or hinder, however, is not the same as actually helping or hindering. A young child who tries to help with the groceries is most likely not actually helping yet. To tell whether someone actually helped (or hindered) requires counterfactuals (95, 126,

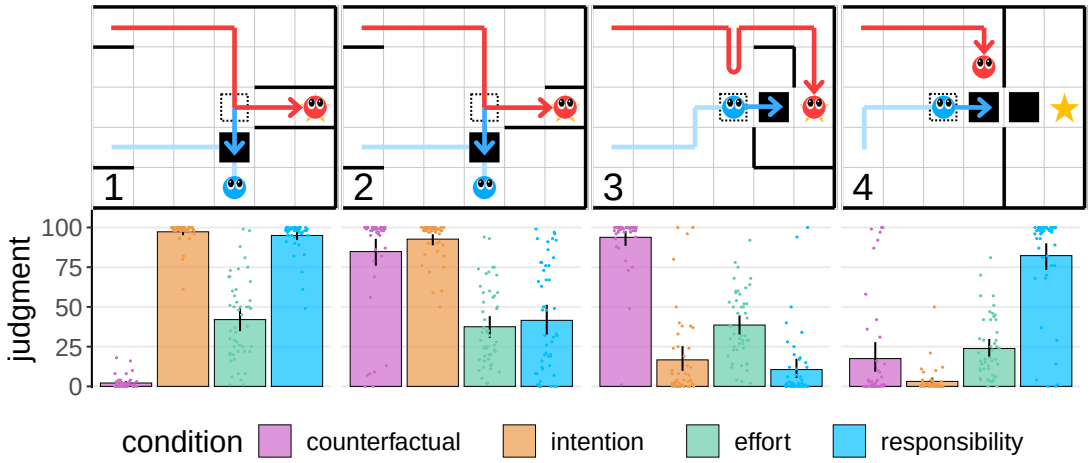


***Figure 4*. The counterfactual simulation model (CSM) predicts responsibility for helping and hindering.** Red's goal is to reach the star in time. Blue's goal is either to help or hinder red. Agents can't move through walls and only blue can pull or push blocks (dashed squares show initial block positions). We asked different groups of participants to judge how likely red would have succeeded without blue (pink), what blue intended to do (orange, low = hinder, high = help), how much effort blue exerted (green), and how responsible blue was for red's success or failure (blue). Note: Bars show means with 95% confidence intervals. Small points are individual participant judgments.

8, 86), and this is where the CSM comes in. Consider the example shown in Figure 1b. The red agent's goal is to reach the star within a given time limit. The blue agent's goal is either to help or hinder red. Neither agent can walk through the barriers and only the blue agent can move the black blocks. In the actual situation, blue pulled the block out of the way and red reached the star in time. To what extent was blue responsible for red's success? When participants answer this question, they care both about what the action reveals about the agent's intention, and what causal role it played (126). For modeling intention inferences, we use the Bayesian theory of mind framework. For modeling causation, we use the CSM. The CSM simulates what would have happened if the blue agent hadn't been present in the scene. The more likely the outcome would have been different from the actual outcome, the more important blue's causal role was (see Figure 1b).

The examples in Figure 4 illustrate how counterfactual simulations and intention inferences jointly affect responsibility judgments. In clips 1 and 2, what red and blue do was identical. However, participants judged blue to be much less responsible for red's success in clip 2 than in clip 1. This is because in clip 2, red would have been able to get to the star even without blue's help because of the opening in the barrier. In

clip 3, participants are sure that red would have reached the star without blue (just like in clip 2). However, this time, blue's responsibility is even lower. This is because, in clip 3, participants inferred that blue's intention was to hinder red, whereas in clip 2, they inferred that blue wanted to help red. Clip 4 also illustrates that intentions matter. Here, it's clear that red would not have succeeded even if blue hadn't been there (because there was already a block in the way), but blue was still viewed as responsible because it wanted to hinder red.

A model that combines intention inferences with counterfactual difference-making explains participants' responsibility judgments well. How much effort an agent exerted mattered less than their intentions. While effort is often highly diagnostic for intentions (107, 12, 127), the two can come apart in our setting, and when they do, intentions matters more. Even though the grid world setup is simple, it supports rich social interactions and inferences. For example, we can implement blue agents that not only have the social goal of helping or hindering, but also presentational goals about the inferences an observer would make from their actions (cf., 128). The actions of the blue agents in clips 3 and 4, for example, are consistent with the goal of wanting to be viewed as a hinderer. Within this framework, we can also predict what kinds of actions someone would take if they wanted to hinder but receive as little responsibility as possible (cf., 16).

When making causal judgments in the physical domain, the counterfactual contrast is often relatively straightforward. For example, when ball A knocked ball B into the gate, people consider what would have happened without ball A. In the social domain, however, it is often not as clear what counterfactual is most relevant (88). For example, instead of considering what would have happened if the blue agent hadn't been there, one could also consider what would have happened if blue had had a different intention. Prior work has shown that intentions matter for how people judge causation in double prevention scenarios like the one shown in Figure 3d (79). When agent B intentionally prevented agent A from preventing outcome E, people are more likely to judge that agent B caused the outcome to happen, compared with when agent B had acted accidentally. Counterfactual theories can account for the fact that intentions matter for causal judgments because intentions make actions more robust causes. Part of what it means to intend an outcome is that one would have adapted one's actions to still bring about the desired outcome if things had played out differently. Research has shown that perceived (counterfactual) robustness affects causal judgments (124, 79, 46, 117, 41).

Instead of imagining that an agent could have had a different mental state, one could

also consider what would have happened if that agent had been replaced by someone else (125). Indeed, the law uses two counterfactual tests to establish causation. The but-for test considers what would have happened but for a defendant's action (73, 111, 55). The reasonable person test considers how a reasonable person would have acted if they had been in the same situation as the defendant (113). An important question for future research is to better understand what counterfactual contrast matters most depending on the situation, and the type of evaluative judgment a person intends to make.

**Box 2: The development of counterfactual thinking**.

Children learn how to reason counterfactually quite late. While earlier work purported to show successful counterfactual reasoning in children as young as two years of age (e.g. 53, 54, 96), later work suggests that these early successes were false alarms (e.g. 7). To demonstrate that a child reasons counterfactually, it's not enough to ask a counterfactual question and get an accurate answer. One needs to show that counterfactual reasoning was required to get the right answer (76). It takes situations where the answers one would get from counterfactual reasoning are different than those from other types of reasoning (such as hypothetical reasoning; see 6, 35). A number of studies have looked at how children reason about causally overdetermined events – where two (or more) events happened that were individually sufficient to make an outcome happen (94, 83, 93). Here, the correct answer is that the outcome would still have happened even if one of the individual events had not happened. Children by the age of 6 or 7 get these questions right (e.g., 83).

One of the challenges with testing how counterfactual reasoning develops is that counterfactual language is complicated. "What would have happened if ..." is a mouthful for an adult, and definitely for a child. To overcome the language barrier, we need to develop experimental paradigms that don't use counterfactual language, but where getting it right requires counterfactual thinking (13, 71). Developing paradigms that don't rely on language would also open the door for studying counterfactual reasoning in non-human primates, and other species (29).

When children get it wrong, there are often many possible reasons for why they failed. They may have not understood the question, construed the counterfactual intervention differently, or had difficulty simulating what the consequences of the counterfactual intervention would have been. To better understand why children get it wrong, we need experimental paradigms that allow us to tease apart these possible

sources of error (see, e.g., 70, 89).

## Concluding remarks

Counterfactual thinking is one of the hallmarks of human intelligence (90, 103). It not only affects how we learn and feel about the world (97), but also how we judge what caused what and who is to blame. I presented the counterfactual simulation model (CSM) – a computational account that captures people's causal and responsibility judgments across a variety of scenarios in the physical and social domain (41, 126).

So far, the CSM has been applied to relatively simple scenarios that feature a small number of objects and agents interacting with one another across a relatively short time frame. In these settings, people can relatively easily imagine how things could have played out differently in relevant counterfactual scenarios. But what about the real world? Does the complexity of our physical and social lives render this approach a non-starter? I don't think so. But generating good explanations of what happened, and why, requires building causal models at the right level of abstraction (101). When we're not able to simulate what would have happened at a detailed lower level, we might succeed at a higher level of abstraction. Counterfactuals are an important tool for breaking down complicated environments into the parts that really matter, and they are now regularly employed to better understand how complex artificial intelligence (AI) systems work (85, 110).

Generative AI is advancing at a rapid pace, building increasingly capable multimodal simulation models of the physical and social world. Equipping these models with the ability for counterfactual simulation will unlock new capabilities (112). Imagine that part of the road accident between the driver and pedestrian from the introduction was caught on camera. Based on this evidence, a generative AI model could build a dynamic 3D reconstruction of what happened and then be asked whether the accident could have been avoided if the driver hadn't been speeding, or if the pedestrian had looked before crossing the street. Of course, going beyond the here and now in such ways will carry important ethical and legal implications (see Outstanding questions; (67)).

**Outstanding questions**

- When do people spontaneously make causal judgments in their everyday lives?
- What functional roles do causal judgments play? Is making causal judgments important for learning?
- Can we develop experimental paradigms that demonstrate counterfactual simulation without the need for counterfactual language (see Box 2), so we can better study the development of counterfactual reasoning in children, primates, and other species?
- What determines what counterfactuals come to mind, and how do we choose the appropriate counterfactual intervention (e.g. removing an action versus replacing a person)?
- How do people carry out counterfactual simulations in their mind? What aspects of the world do they choose to simulate, what aspects do they ignore?
- When explaining other's behavior one can consider counterfactuals at different levels of abstraction (e.g. over actions, mental states, or traits). How do people choose the "right" level of abstraction?
- Can CSMs help clarify conceptual distinctions between various social and moral judgments (e.g. causation, responsibility, blame, permissibility)?
- How can CSMs scale up to deal with more complex domains?
- How can CSMs best interface with multi-modal generative AI models to produce human-understandable explanations of what happened and why?

**Acknowledgments**

I thank David Rose, Philipp Fränken, Thomas Icard, the members of the Causality in Cognition Lab, and the SAME writing group for feedback and discussion. I thank Herbert Clark for detailed feedback on the manuscript. I was supported by a grant from the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

References

[1] 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.

[2] Jamie Amemiya, Gail D. Heyman, and Tobias Gerstenberg. Children use disagreement to infer what happened. *PsyArXiv*, 2023. URL `https://psyarxiv.com/y79sd/`.

[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. doi: $10.1038/s41562\text{-}017\text{-}0064$. URL `https://doi.org/10.1038%2Fs41562-017-0064`.

[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] Sarah R. Beck. Why what is counterfactual really matters: A response to Weisberg and Gopnik (2013). *Cognitive Science*, 40(1):253–256, 2015. doi: $10.1111/\mathrm{cogs}.12235$. URL `https://doi.org/10.1111%2Fcogs.12235`.

[7] Sarah R Beck and Carlie Guthrie. Almost thinking counterfactually: Children's understanding of close counterfactuals. *Child development*, 82(4):1189–1198, 2011.

[8] Sander Beckers, Hana Chockler, and Joseph Halpern. A causal analysis of harm. *Advances in Neural Information Processing Systems*, 35:2365–2376, 2022.

[9] H. Beebee. Causing and nothingness. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and counterfactuals*, pages 291–308. MIT Press Cambridge, MA, 2004.

[10] Aaron Beller and Tobias Gerstenberg. A counterfactual simulation model of causal language. *PsyArXiv*, 2023. URL `https://psyarxiv.com/xv8hf`.

[11] Aaron Beller, Erin Bennett, and Tobias Gerstenberg. The language of causation. In S. Denison, M. Mack, Y. Xu, and B. C. Armstrong, editors, *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 3133–3139. Cognitive Science Society, 2020.

[12] Yochanan E Bigman and Maya Tamir. The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, 145(12):1654, 2016.

[13] Sophie Bridgers, Chuyi Yang, Tobias Gerstenberg, and Hyo Gweon. Whom will granny thank? thinking about what could have been informs children's inferences about relative helpfulness. In S. Denison., M. Mack, Y. Xu, and B.C. Armstrong, editors, *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 2446–2452, 2020.

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

[15] Angela Cao, Atticus Geiger, Elisa Kreiss, Thomas Icard, and Tobias Gerstenberg. A semantics for causing, enabling, and preventing verbs using structural causal models. In Micah B. Goldwater, Florencia Anggoro, Brett Hayes, and Desmond C Ong, editors, *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, pages 2947–2954, 2023.

[16] Kartik Chandra, Tzu-Mao Li, Josh Tenenbaum, and Jonathan Ragan-Kelley. Acting as inverse inverse planning. *arXiv preprint arXiv:2305.16913*, 2023.

[17] Nick Chater and Mike Oaksford. Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, 37(6):1171–1191, 2013.

[18] P. W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997.

[19] P. W. Cheng and L. R. Novick. Causes versus enabling conditions. *Cognition*, 40: 83–120, 1991.

[20] Hana Chockler and Joseph Y. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22(1):93–115, 2004.

[21] Eve V Clark and Herbert H Clark. When nouns surface as verbs. *Language*, pages 767–811, 1979.

[22] Kenneth James Williams Craik. *The nature of explanation*. Cambridge University Press, Cambridge, UK, 1943.

[23] F. Cushman. Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2):353–380, 2008.

[24] David Danks. Singular causation. In Michael Waldmannn, editor, *The Oxford Handbook of Causal Reasoning*, pages 201–215. Oxford University Press, 2017.

[25] Julian De Freitas and George A Alvarez. Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition*, 178:133–146, 2018.

[26] Judith Degen. The rational speech act framework. *Annual Review of Linguistics*, 9:519–540, 2023.

[27] D. C. Dennett. *The intentional stance*. MIT Press, Cambridge, MA, 1987.

[28] P. Dowe. *Physical Causation*. Cambridge University Press, Cambridge, England, 2000.

[29] Jan M Engelmann, Christoph J Völter, Cathal O'Madagain, Marina Proft, Daniel BM Haun, Hannes Rakoczy, and Esther Herrmann. Chimpanzees consider alternative possibilities. *Current Biology*, 31(20):1377–1378, 2021.

[30] Florian Engl. A theory of causal responsibility attribution. 2022. URL `https://dx.doi.org/10.2139/ssrn.2932769`.

[31] Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

[32] Max Frisch. *Biography: A Game (new Version, 1984)*. Seagull Books Pvt Ltd, 2010.

[33] T. Gerstenberg and N. D. Goodman. Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, and R. P. Cooper, editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1590–1595. Austin, TX: Cognitive Science Society, 2012.

[34] T. Gerstenberg and D. A. Lagnado. Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, 115(1):166–171, 2010.

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

[36] Tobias Gerstenberg and Thomas F. Icard. Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, 149(3):599–607, 2020.

[37] Tobias Gerstenberg and Simon Stephan. A counterfactual simulation model of causation by omission. *Cognition*, 216:104842, 2021. URL `https://psyarxiv.com/wmh4c/`.

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

[39] 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. doi: 10.1177/0956797617713053. URL https://doi.org/10.1177%2F0956797617713053.

[40] 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. ISSN 00100277. doi: 10.1016/j.cognition.2018.03.019.

[41] 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.

[42] Tobias Gerstenberg, David A Lagnado, et al. Making a positive difference: Criticality in groups. *Cognition*, 238:105499, 2023.

[43] E. Goldvarg and P. N. Johnson-Laird. Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, 25(4):565–610, 2001.

[44] N. D. Goodman, J. B. Tenenbaum, and T. Gerstenberg. Concepts in a probabilistic language of thought. In Eric Margolis and Stephen Lawrence, editors, *The Conceptual Mind: New Directions in the Study of Concepts*, pages 623–653. MIT Press, 2015.

[45] Noah D. Goodman, Vikash Mansinghka, Daniel M. Roy, Kallista A. Bonawitz, and Joshua B. Tenenbaum. Church: a language for generative models. *CoRR*, abs/1206.3255, 2014. URL http://arxiv.org/abs/1206.3255.

[46] Guy Grinfeld, David Lagnado, Tobias Gerstenberg, James F. Woodward, and Marius Usher. Causal responsibility and robust causation. *Frontiers in Psychology*, 11: 1069, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.01069. URL https://www.frontiersin.org/article/10.3389/fpsyg.2020.01069.

[47] Hyowon Gweon. Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, 25(10):896–910, 2021.

[48] 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.

[49] N. Hall. Two concepts of causation. In J. Collins, N. Hall, and L. A. Paul, editors, *Causation and Counterfactuals*. MIT Press, 2004.

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

[51] Joseph Y Halpern. *Actual causality*. MIT Press, 2016.

[52] Joseph Y Halpern and Christopher Hitchcock. Graded causation and defaults. *British Journal for the Philosophy of Science*, 66:413–457, 2015.

[53] Paul L Harris, Tim German, and Patrick Mills. Children's use of counterfactual thinking in causal reasoning. *Cognition*, 61(3):233–259, 1996.

[54] PL Harris. On realizing what might have happened instead. *Polish Quarterly of Developmental Psychology*, 3:161–176, 1997.

[55] H. L. A. Hart and T. Honoré. *Causation in the law*. Oxford University Press, New York, 1959/1985.

[56] Paul Henne and Kevin O'Neill. Double prevention, causal judgments, and counterfactuals. *Cognitive Science*, 46(5):e13127, 2022.

[57] Paul Henne, Ángel Pinillos, and Felipe De Brigard. Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, 95(2):270–283, 2017.

[58] 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. ISSN 00100277. doi: 10.1016/j.cognition.2019.05.006.

[59] Denis J. Hilton. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4):273–308, 1996. ISSN 1464-0708. doi: 10.1080/135467896394447. URL http://dx.doi.org/10.1080/135467896394447.

[60] Thomas F. Icard, Jonathan F. Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017. doi: 10.1016/j.cognition.2017.01.010. URL https://doi.org/10.1016%2Fj.cognition.2017.01.010.

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

[62] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[63] 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. doi: 10.1016/j.tics.2016.08.007. URL https://doi.org/10.1016%2Fj.tics.2016.08.007.

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

[65] 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.

[66] Sangeet Khemlani, Christina Wasylyshyn, Gordon Briggs, and Paul Bello. Mental models and omissive causation. *Memory & Cognition*, 46(8):1344–1359, 2018.

[67] Lara Kirfel, Robert J. MacCoun, Thomas Icard, and Tobias Gerstenberg. Anticipating the risks and benefits of counterfactual world simulation models. In *AI Meets Moral Philosophy and Moral Psychology Workshop (NeurIPS 2023)*, 2023.

[68] M. Kleiman-Weiner, T. Gerstenberg, S. Levine, and J. B. Tenenbaum. Inference of intention and permissibility in moral decision making. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, Jennings Matlock, T., C. D., and P. P. Maglio, editors, *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 1123–1128, Austin, TX, 2015. Cognitive Science Society.

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

[70] Jonathan F. Kominsky, Tobias Gerstenberg, Madeline Pelz, Mark Sheskin, Henrik Singmann, Laura Schulz, and Frank C. Keil. The trajectory of counterfactual simulation in development. *Developmental Psychology*, 57(2):253–268, 2021. ISSN 1939-0599, 0012-1649. doi: $10.1037/\mathrm{dev}0001140$. URL `http://doi.apa.org/getdoi.cfm?doi=10.1037/dev0001140`.

[71] Karla Koskuba, Tobias Gerstenberg, Hannah Gordon, David A. Lagnado, and Anne Schlottmann. What's fair? how children assign reward to members of teams with differing causal structures. *Cognition*, 177:234–248, 2018.

[72] James R. Kubricht, Keith J. Holyoak, and Hongjing Lu. Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21(10):749–759, 2017. doi: $10.1016/\mathrm{j.tics}.2017.06.002$. URL `https://doi.org/10.1016%2Fj.tics.2017.06.002`.

[73] D. A. Lagnado and T. Gerstenberg. Causation in legal and moral reasoning. In Michael Waldmann, editor, *Oxford Handbook of Causal Reasoning*, pages 565–602. Oxford University Press, 2017.

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

[75] 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.

[76] Brian Leahy, Eva Rafetseder, and Josef Perner. Basic conditional reasoning: How children mimic counterfactual reasoning. *Studia logica*, 102:793–810, 2014.

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

[78] Jonathan Livengood and Edouard Machery. The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, 31(1):107–127, 2007.

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

[80] Ethan Ludwin-Peery, Neil R Bramley, Ernest Davis, and Todd M Gureckis. Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, 127:101396, 2021.

[81] Bertram F. Malle, Steve Guglielmo, and Andrew E. Monroe. A theory of blame. *Psychological Inquiry*, 25(2):147–186, 2014. doi: 10.1080/1047840x.2014.877340. URL http://dx.doi.org/10.1080/1047840x.2014.877340.

[82] D. R. Mandel. Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General*, 132(3):419–434, 2003.

[83] Teresa McCormack, Margaret Ho, Charlene Gribben, Eimear O'Connor, and Christoph Hoerl. The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, 45:1–9, 2018. doi: 10.1016/j.cogdev.2017.10.001. URL https://doi.org/10.1016%2Fj.cogdev.2017.10.001.

[84] S. McGrath. Causation by omission: A dilemma. *Philosophical Studies*, 123(1): 125–148, 2005.

[85] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007. URL http://dx.doi.org/10.1016/j.artint.2018.07.007.

[86] Scott Mueller and Judea Pearl. Personalized decision making–a conceptual introduction. *Journal of Causal Inference*, 11(1):20220050, 2023.

[87] Prerna Nadathur and Sven Lauer. Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: A journal of general linguistics*, 5(1), 2020.

[88] Laura Niemi, Joshua Hartshorne, Tobias Gerstenberg, Matthew Stanley, and Liane Young. Moral values reveal the causality implicit in verb meaning. *Cognitive Science*, 44(6), 2020. ISSN 0364-0213, 1551-6709. doi: 10.1111/cogs.12838.

[89] Joseph Outa, Xi Jia Zhou, Hyowon Gweon, and Tobias Gerstenberg. Stop, children what's that sound? multi-modal inference through mental simulation. In Jennifer Culbertson, Andrew Perfors, Hugh Rabagliati, and Veronica Ramenzoni, editors, *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, pages 1359–1366. Cognitive Science Society, 2022.

[90] Judea Pearl and Dana Mackenzie. *The book of why: The new science of cause and effect*. Basic Books, New York, 2018.

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

[92] Tadeg Quillien and Christopher G Lucas. Counterfactuals and the logic of causal selection. *Psychological Review*, 2023.

[93] Eva Rafetseder and Josef Perner. Belief and counterfactuality: A teleological theory of belief attribution. *Zeitschrift für Psychologie*, 226(2):110–121, 2018. ISSN 2190-8370, 2151-2604. doi: $10.1027/2151\text{-}2604/\mathrm{a}000327$.

[94] Eva Rafetseder, Maria Schwitalla, and Josef Perner. Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, 114(3): 389–404, 2013. ISSN 00220965. doi: $10.1016/\mathrm{j.jecp}.2012.10.010$.

[95] Jonathan Richens, Rory Beard, and Daniel H Thompson. Counterfactual harm. *Advances in Neural Information Processing Systems*, 35:36350–36365, 2022.

[96] Kevin J Riggs, Donald M Peterson, Elizabeth J Robinson, and Peter Mitchell. Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, 13(1):73–90, 1998.

[97] N. J. Roese. Counterfactual thinking. *Psychological Bulletin*, 121(1):133–148, 1997.

[98] David Rose, Eric Sievers, and Shaun Nichols. Cause and burn. *Cognition*, 207 (104517), 2021.

[99] W. C. Salmon. Causality without counterfactuals. *Philosophy of Science*, 61(2): 297–312, 1994.

[100] K. G. Shaver. *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer-Verlag, New York, 1985.

[101] Steven M Shin and Tobias Gerstenberg. Learning what matters: Causal abstraction in human inference. In Micah B. Goldwater, Florencia Anggoro, Brett Hayes, and Desmond C Ong, editors, *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, pages 115–122, 2023. URL `https://psyarxiv.com/br2vz`.

[102] Tianmin Shu, Marta Kryven, Tomer D Ullman, and Joshua B Tenenbaum. Adventures in flatland: Perceiving social interactions under physical dynamics. In S. Denison, M. Mack, Y. Xu, and B. C. Armstrong, editors, *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 2901–2907, 2020.

[103] S. A. Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, USA, 2005.

[104] S. A. Sloman, A. K. Barbey, and J. M. Hotaling. A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, 33(1):21–50, 2009.

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

[106] Kevin A. Smith, Jessica B. Hamrick, Adam N. Sanborn, Peter W. Battaglia, Tobias Gerstenberg, Tomer D. Ullman, and Joshua B. Tenenbaum. Probabilistic models of physical reasoning. In Tom L. Griffiths, Nick Chater, and Joshua B. Tenenbaum, editors, *Reverse engineering the mind: Probabilistic models of cognition*. MIT Press, in press.

[107] Felix A Sosa, Tomer Ullman, Joshua B Tenenbaum, Samuel J Gershman, and Tobias Gerstenberg. Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, 217:104890, 2021.

[108] Simon Stephan and Michael R. Waldmann. Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, 10(1):242–257, 2018. doi: $10.1111/\mathrm{tops}.12309$. URL `https://doi.org/10.1111%2Ftops.12309`.

[109] Simon Stephan, Ralf Mayrhofer, and Michael R Waldmann. Time and singular causation: A computational model. *Cognitive Science*, 44(7):e12871, 2020.

[110] Ilia Stepin, Jose M Alonso, Alejandro Catala, and Martín Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[111] Andrew Summers. Common-sense causation in the law. *Oxford Journal of Legal Studies*, 38(4):793–821, 2018. ISSN 1464-3820. doi: $10.1093/\mathrm{ojls/gqy028}$. URL `http://dx.doi.org/10.1093/ojls/gqy028`.

[112] Zenna Tavares, James Koppel, Xin Zhang, Ria Das, and Armando Solar-Lezama. A language for counterfactual generative models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10173–10182. PMLR, 18–24 Jul 2021.

[113] Kevin P Tobia. How people judge what is reasonable. *Alabama Law Review*, 70: 293–359, 2018.

[114] 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.

[115] Tomer D Ullman and Joshua B Tenenbaum. Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, 2:533–558, 2020.

[116] 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. doi: 10.1016/j.tics.2017.05.012. URL `https://doi.org/10.1016%2Fj.tics.2017.05.012`.

[117] Nadya Vasilyeva, Thomas Blanchard, and Tania Lombrozo. Stable causal relationships are better causal relationships. *Cognitive Science*, 42(4):1265–1296, 2018. ISSN 0364-0213. doi: 10.1111/cogs.12605. URL `http://dx.doi.org/10.1111/cogs.12605`.

[118] Michael R Waldmann, Jonas Nagel, and Alex Wiegmann. Moral judgment. In *The Oxford handbook of Thinking and Reasoning*, pages 364–389. Oxford University Press, New York, 2012.

[119] C. R. Walsh and S. A. Sloman. The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, 26(1):21–52, 2011.

[120] B. Weiner. *Judgments of responsibility: A foundation for a theory of social conduct*. The Guilford Press, New York, 1995.

[121] P. Wolff. Representing causation. *Journal of Experimental Psychology: General*, 136(1):82–111, 2007.

[122] P. Wolff, A. K. Barbey, and M. Hausknecht. For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, 139(2):191–221, 2010.

[123] J. Woodward. *Making things happen: A theory of causal explanation*. Oxford University Press, Oxford, England, 2003.

[124] J. Woodward. Sensitive and insensitive causation. *The Philosophical Review*, 115 (1):1–50, 2006.

[125] Sarah A Wu and Tobias Gerstenberg. If not me, then who? Responsibility and replacement. *Cognition*, 242:105646, 2024.

[126] Sarah A. Wu, Shruti Sridhar, and Tobias Gerstenberg. A computational model of responsibility judgments from counterfactual simulations and intention inferences. In Micah B. Goldwater, Florencia Anggoro, Brett Hayes, and Desmond C Ong, editors, *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, pages 3375–3382, 2023. URL `https://psyarxiv.com/uwdbr/`.

[127] Yang Xiang, Jenna Landy, Fiery Cushman, Natalia Vélez, and Samuel J Gershman. Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, 2023.

[128] Erica J Yoon, Michael Henry Tessler, Noah D Goodman, and Michael C Frank. Polite speech emerges from competing social goals. *Open Mind*, 4:71–87, 2020.

[129] Liang Zhou, Kevin A. Smith, Joshua B. Tenenbaum, and Tobias Gerstenberg. Mental jenga: A counterfactual simulation model of causal judgments about physical support. *Journal of Experimental Psychology: General*, 152(8):2237–2269, 2023.

[130] R. Zultan, T. Gerstenberg, and D. A. Lagnado. Finding fault: Counterfactuals and causality in group attributions. *Cognition*, 125(3):429–440, 2012.