What would have happened? Counterfactuals, hypotheticals, and causal judgments

Tobias Gerstenberg
Stanford University

Abstract

How do people make causal judgments? In this paper, I show that counterfactual simulations are necessary for explaining causal judgments about events, and that hypotheticals don't suffice. In two experiments, participants viewed video clips of dynamic interactions between billiard balls. In Experiment 1, participants either made hypothetical judgments about whether ball B *would go* through the gate if ball A weren't present in the scene, or counterfactual judgments about whether ball B *would have gone* through the gate if ball A hadn't been present. Because the clips featured a block in front of the gate that sometimes moved and sometimes stayed put, hypothetical and counterfactual judgments came apart. A computational model that evaluates hypotheticals and counterfactuals by running noisy physical simulations accurately captured participants' judgments. In Experiment 2, participants judged whether ball A caused ball B to go through the gate. The results showed a tight fit between counterfactual and causal judgments, whereas hypotheticals didn't predict causal judgments. I discuss the implications of this work for theories of causality, and for studying the development of counterfactual thinking in children.

Keywords: causality; counterfactual; hypothetical; conditional; mental simulation; intuitive physics.

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

Introduction

How do people make causal judgments? Consider the diagram shown in Figure 1a. Ball A and ball B enter the scene from the right, collide with one another, and ball B goes through the gate. Did ball A cause ball B to go through the gate? Intuitively, the answer is 'yes'. But why?

Reaching causal verdicts about scenes like this one requires going beyond what actually happened, and considering what would have happened in a relevant counterfactual situation (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Kahneman & Tversky, 1982; Sloman & Lagnado, 2015). If ball A hadn't been present in the scene then ball B wouldn't have gone through the gate. The fact that this counterfactual is true suggests that ball A caused ball B to go through the gate. By the same logic, in Figure 1b, ball A didn't cause ball B to go through the gate. Here, ball B would have gone through the gate even if ball A hadn't been present in the scene. Gerstenberg, Peterson, Goodman, Lagnado, and Tenenbaum (2017) show that a model based on counterfactual simulation accurately captures people's causal judgments about physical events like this one (see also Gerstenberg et al., 2021; Gerstenberg & Stephan, 2021). But are counterfactuals really necessary, or might it be possible to explain causal judgments differently? One such possibility is that another kind of cognitive operation may suffice: hypothetical simulation. But what's the difference between counterfactual and hypothetical simulation?

A *counterfactual simulation* involves observing what actually happened, mentally travelling back in time to imagine a change to what actually happened, and then simulating how this alternative possibility would have played out. If the outcome in the counterfactual situation would have been different from what actually happened then the event of interest caused the outcome. In contrast, a *hypothetical simulation* involves imagining a possible future. This doesn't require going back in time and mentally changing something that already happened. Instead, one considers a possible change in the future. While the

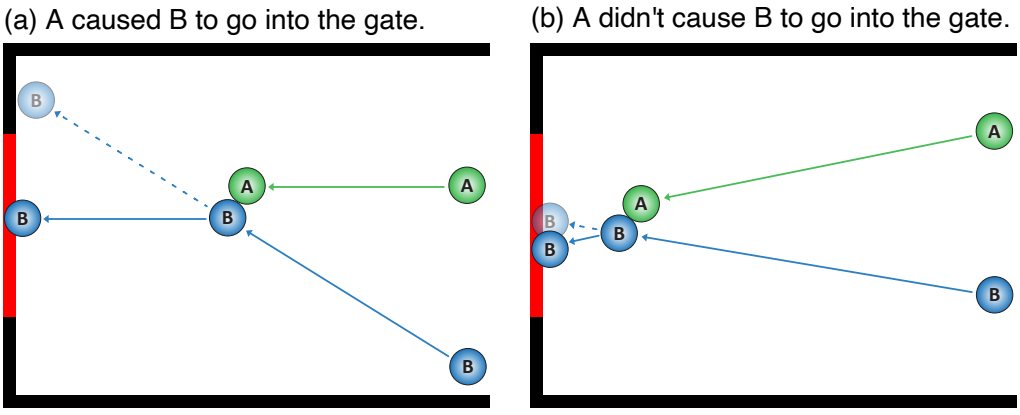

Figure 1. Two diagrams illustrating the difference between a situation in which (a) ball A caused ball B to go through the gate, and (b) one in which ball A didn't cause ball B to go through the gate. In (a) ball B would have missed the gate if ball A hadn't been present in the scene. In (b) ball B would have gone through the gate even if ball A hadn't been present.

counterfactual asks whether ball B *would have gone* through the gate if ball A *hadn't been there*, the hypothetical asks whether ball B *would go* through the gate if ball A *weren't there*. So, counterfactuals and hypotheticals differ in whether the mind travels to the past or to the future.

For example, when judging causation in Figure 1a, as the balls enter the scene an observer may consider a *hypothetical* simulation of where ball B would go if ball A weren't present in the scene, and then compare what actually happened to the outcome of this future hypothetical. In fact, for the clips shown in Figure 1, the hypothetical probability (would B go through the gate if ball A weren't there) and the counterfactual probability (would B have gone through the gate if ball A hadn't been there) are the same. These cases can't distinguish between what kind of mental time travel is involved in making causal judgments. We need new evidence to determine whether counterfactuals are necessary for explaining causal judgments, or whether hypotheticals suffice.[1]

In this paper, I present new evidence that bears on the question of what kind of mental simulation is involved when people make causal judgments. First, I will clarify the conceptual distinction between conditionals, hypotheticals, and counterfactuals using the formal framework developed by Pearl (2000). I will then discuss prior research focusing on the role that counterfactuals play in theories of causal judgment. The empirical evidence so far doesn't conclusively show that people engage in counterfactual simulation when making causal judgments. I will present such evidence. I develop a physical simulation model that generates both hypothetical and counterfactual simulations, and test the model in two experiments. The experiments feature a set of physical scenarios in which the outcomes of hypothetical and counterfactual simulations differ. Experiment 1 shows that the model accurately captures participants' hypothetical and counterfactual judgments. Experiment 2 tests whether causal judgments are better explained by hypotheticals or counterfactuals. The results clearly support the counterfactual account. I discuss the implications of these findings for theories of causality, and for psychological research into the development of counterfactual reasoning.

Counterfactuals versus hypotheticals: same, same but different

Counterfactuals and hypotheticals are both thoughts about possibilities. The key difference is *when* a change to actuality is imagined to take place. Hypotheticals are thoughts about *changes that lie in the future*. For example, as ball A and ball B enter the scene in Figure 2a and before they are about to collide with one another, one might wonder whether ball B would go through the gate if ball A was removed from the scene. Imagining such a future hypothetical, an observer mentally removes ball A from the scene and simulates what path ball B would take. Hypotheticals are essential for decision-making and planning (Sloman & Hagmayer, 2006). Making good decisions requires evaluating the likely consequences of different hypothetical actions and then choosing the action with the highest

[1]Hypotheticals are also sometimes expressed using indicative language (Conditional I: "If this thing happens, that thing will happen.") rather than subjunctive language (Conditional II: "If this thing happened, that thing would happen."). I prefer to use the subjunctive form because I think it helps to differentiate between conditionals that are based on observations, those that are based on (hypothetical) interventions, and those that are based on counterfactuals (Conditional III: "If this thing had happened, that thing would have happened".).

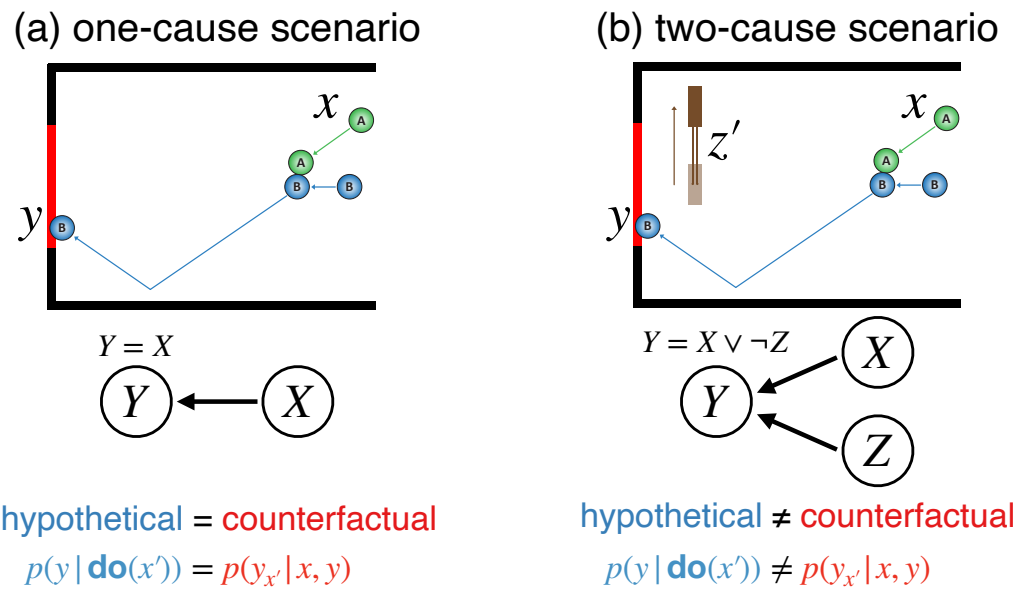

Figure 2. Illustration of the (a) one-cause scenario, and (b) two-cause scenario. The images show what happened in each scenario. The diagrams and structural equations below capture the causal dependence between the variables that represent the relevant events (X = whether ball A is present or absent, Y = whether ball B goes through the gate or nor, Z = whether the block ends up in front of the gate or not). In the *one-cause scenario*, the hypothetical probability ("Would ball B go through the gate if ball A weren't there?") and the counterfactual probability ("Would ball B have gone through the gate if ball A hadn't there?") are the same. In the *two-cause scenario*, the hypothetical and counterfactual probability come apart because of the uncertainty about whether or not the block will move. The hypothetical probability is determined at the beginning of the clip (at which point it's unclear whether or not the block will move), whereas the counterfactual probability is determined after the clip has played (at which point it is now clear whether or not the block moved).

expected utility.

Counterfactuals are thoughts about *changes that lie in the past*. For example, after ball B ended up going through the gate in Figure 2a, one might wonder whether ball B would still have gone through the gate even if ball A hadn't been present. Here, the observer first takes in everything that happened, then goes back in time, and mentally simulates how things would have played out if the change of interest had taken place.

Note that for the scene shown in Figure 2a, the hypothetical and counterfactual outcome are the same. Where ball B would go if ball A wasn't present (the hypothetical) is the same as where ball B would have gone if ball A hadn't been present (the counterfactual). This is the case because there are no other factors that influence the outcome (beyond the presence or absence of ball A) about which an observer may have some degree of uncertainty (e.g. other balls that could enter at a later point, the possibility that the gate may close, ...). However, as we will see soon, hypotheticals and counterfactuals can come apart.

A hierarchy of causal concepts. Formal theories of causality make a conceptual distinction between hypotheticals and counterfactuals. Pearl and Mackenzie (2018) propose a metaphorical ladder of causation where each rung represents what kinds of causal questions can be answered (Pearl, 2000, 2019). Let's consider a simple setting with two binary variables $X \in \{x, x'\}$ (the candidate cause), and $Y \in \{y, y'\}$ (the candidate effect), where x (or y) indicates that the event of interest happened, and x' (or y') indicates that it didn't happen. Table 1 summarizes the three levels of the causal hierarchy (see also Bareinboim, Correa, Ibeling, & Icard, 2020).

On level I of the ladder, one can answer conditional questions such as how likely y will happen if x happens, $p(y|x)$. On level II, one can answer hypothetical questions such as how likely y would happen if x were made true, $p(y|\text{do}(x))$. The formally defined concept of an intervention, do(), distinguishes causal from merely correlational relationships (Pearl, 2000). Intuitively, when X causes Y, intervening on X increases the probability that Y will happen (i.e. $p(y|\text{do}(x)) > p(y)$). Whereas when X and Y are merely correlated, it's possible that this correlation is due to another factor, such as a common cause C that brings about both X and Y (in this case, $p(y|\text{do}(x)) = p(y)$). Finally, on level III, one can answer counterfactual questions such as whether y would have happened if x had been made true, given that in fact neither x nor y happened, $p(y_x|x', y')$. Answering counterfactual questions requires a combination of conditioning on what actually happened, and then (mentally) changing an event that already happened, to see whether things would have played out differently.

As we have seen above, hypothetical and counterfactual probabilities do not always come apart. Let's say that in Figure 2a, X denotes whether or not ball A was present in

Table 1
The causal hierarchy adapted from Pearl (2019). On level I, one can only answer questions about probabilistic dependence. On level II, one can distinguish genuine causation from mere correlation. On level III, one can answer questions about why a particular event of interest happened.

Level	Concept	Expression	Activity	Question	Example
I	Observation/ Prediction	$p(y\|x)$	Seeing	How does x change my belief in y?	Would the grass be dry if we *found* the sprinkler off?
II	Intervention/ Hypothetical	$p(y\|\text{do}(x))$	Doing	Would y happen if I did x?	Would the the grass be dry if we *made sure* that the sprinkler was off?
III	Counterfactual	$p(y_x\|x', y')$	Explaining	Would y have happened instead of y', if I had done x instead of x'?	Would the grass have been dry if the sprinkler had been off, given that the grass is wet and the sprinkler on?

the scene with $X \in \{x = \text{ball A is present}, x' = \text{ball A is absent}\}$, and Y denotes whether or not ball B goes through the gate with $Y \in \{y = \text{ball B goes through the gate}, y' = \text{ball B doesn't go through the gate}\}$. In this simple setting, the hypothetical probability $p(y|\text{do}(x'))$ is the same as the counterfactual probability $p(y_{x'}|x, y)$. The probability that ball B would go through the gate if ball A weren't there is the same as the probability that ball B would have gone through if ball A hadn't been there. This is the case because there aren't any other factors that influence ball B's going through the gate except for ball A. In order for hypothetical and counterfactual probabilities to come apart, we need to go beyond such simple one-cause scenarios.

Figure 2b shows a situation in which ball A and ball B interact with one another in the same way as they did in Figure 2a. However, this time there is another object in the scene that affects the outcome: a block in front of the gate that sometimes moves and sometimes stays put. I'll use the variable Z for the block with $Z \in \{z = \text{the block is in front of the gate}, z' = \text{the block is not in front of the gate}\}$. Let's assume that the block has a 50% chance of moving $p(Z = z) = 0.5$. In this scenario, the hypothetical probability and the counterfactual probability come apart. The hypothetical probability that ball B would go through the gate if ball A weren't present is $p(y|\text{do}(x')) \approx 0.5$. Ball B would only go through the gate (in ball A's absence) if the block moved, but whether or not this will happen is unclear at the beginning of the clip. The counterfactual probability that ball B would have gone through the gate if ball A hadn't been present is $p(y_{x'}|x, z', y) \approx 1$. This is because when considering the counterfactual, we condition on what actually happened: ball A was present (x), the block wasn't in front of the gate (z'), and ball B went through the gate (y). Because the counterfactual intervention on A's presence doesn't affect whether or not the block moves (there is no causal link from X to Z), it's clear that the block would still have been out of the way in the counterfactual situation in which ball A hadn't been present, and that ball B would still have gone through the gate in that case.

The sprinkler example. As another illustration of how conditioning on observations, interventions, or counterfactuals comes apart, consider the example shown in Figure 3 (adapted from Pearl, 2000). Picture yourself in sunny California wondering whether it was the sprinkler ($S \in \{s = \text{sprinkler on}, s' = \text{sprinkler off}\}$) or the rain ($R \in \{r = \text{rain present}, r' = \text{rain absent}\}$) that caused the grass on your lawn to be wet ($W \in \{w = \text{grass is wet}, w' = \text{grass is dry}\}$). The clouds ($C \in \{c = \text{clouds present}, c' = \text{clouds absent}\}$) cause the rain, and they prevent the sprinkler from running (it's one of these Silicon valley smart sprinklers that only runs if there aren't any clouds). The grass is wet if the sprinkler is on, if it rains, or if both are the case (Figure 3a). Somewhat unrealistically, there is a 50% chance on any given day that there are clouds, $p(C = c) = 0.5$.

On Monday morning, you go outside and you see that there are no clouds, that the sprinkler is on, that there is no rain, and that the grass is wet (Figure 3b). You wonder, did the sprinkler cause the grass to be wet, $p(s \rightarrow w)$? Intuitively, the answer is 'yes', of course. After all, the sprinkler was on, and there was no rain. But what verdict would we reach based on the three different levels in the causal hierarchy?

To answer the question of whether the sprinkler caused the grass to be wet, we want to test whether the grass would have been dry if the sprinkler had been off. On level I, we can only condition on observations (Figure 3c). Observing that the sprinkler is off licenses the

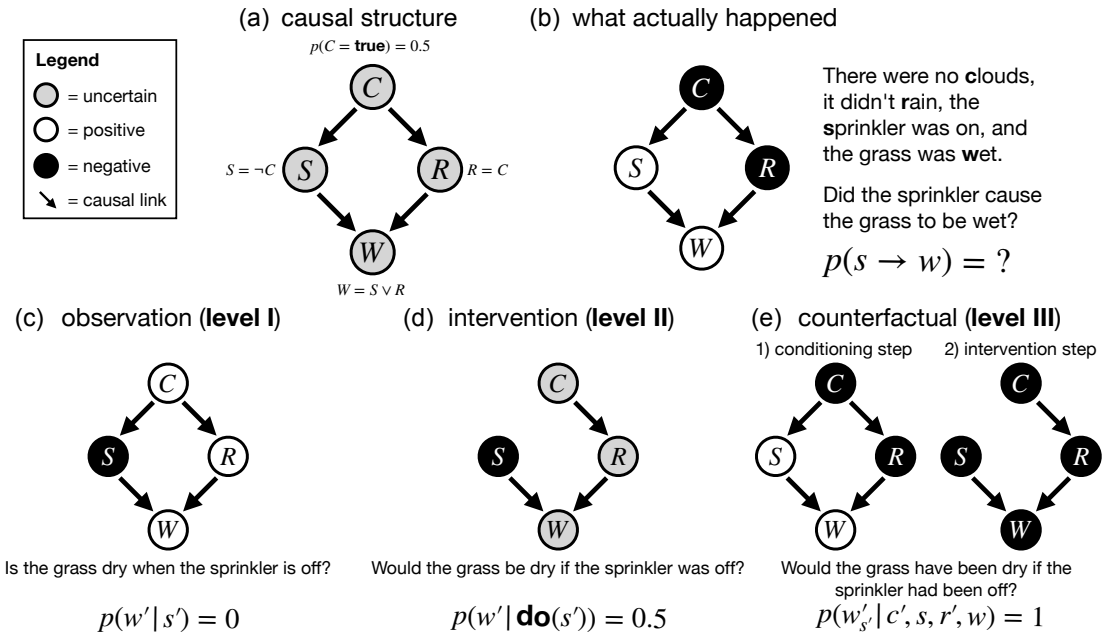

Figure 3. Diagrammatic illustration of the difference between making inferences based on observations, interventions, and counterfactuals. a) The causal structure of the setting. b) What actually happened. c) *Observation*: What can be inferred from observing that the sprinkler is off? d) *Intervention*: What can be inferred from intervening to turn the sprinkler off? d) *Counterfactual*: What can be inferred from first observing what actually happened, and then intervening to turn the sprinkler off? Here, only the counterfactual level yields the intuitively correct response that the sprinkler caused the grass to be wet in the actual situation.

diagnostic inference that there must be clouds, which in turn means that it rained, which in turn means that the grass is wet. So, on this level, the grass would not be dry if one observed the sprinkler to be off, $p(w'|s') = 0$. On level II, we condition on hypothetically intervening on the scene (Figure 3d). Intervening to turn the sprinkler off breaks the causal link between clouds (C) and sprinkler (S). Intervening on a variable removes all the incoming links into that variable (and thereby breaks any diagnostic inferences from the intervened-on variable to its parents). There is a 50% chance that the grass would be dry if we intervened to turn the sprinkler off because it now depends on whether or not there would be clouds, $p(w'|\text{do}(s')) = 0.5$. On level III, we first condition on what actually happened, and then consider a counterfactual intervention that would have turned the sprinkler off (Figure 3e). This yields the inference that the grass would have been dry had the sprinkler been turned off, $p(w'_{s'}|c', s, r', w) = 1$. So, only on the counterfactual level do we get the intuitive verdict that it was sprinkler that caused the grass to be wet.

Prior work

There is a vast literature on how conditional and counterfactual reasoning relates to causality in philosophy (e.g. Adams, 1965; Edgington, 1995; Lewis, 1976; Stalnaker, 1970),

linguistics (Ciardelli, Zhang, & Champollion, 2018; Kaufmann, 2013; Kratzer, 1981; Lassiter, 2017a, 2017b; Schulz, 2011), and psychology (e.g. Byrne, 2016, 2005; Cheng, 1997; Douven & Verbrugge, 2010; Oaksford & Chater, 2007; Over & Evans, 2003; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; van Rooij & Schulz, 2019). Here, I will focus on work in psychology that has directly been inspired by Pearl's (2000) formal modeling framework (for an overview, see Sloman & Lagnado, 2015).

A lot of work has shown that people differentiate between levels I and II of the causal hierarchy (see Table 1). People are sensitive to the different inferences that are licensed based on 'seeing' versus 'doing' (Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Gopnik et al., 2004; McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Meder, Gerstenberg, Hagmayer, & Waldmann, 2010; Sloman & Hagmayer, 2006; Sloman & Lagnado, 2005; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). For example, whereas observing an effect makes it more likely that a cause was present, intervening on an effect blocks the diagnostic inference about the likelihood of its cause.

Psychologists have also studied whether the way in which people reason about counterfactuals accords with Pearl's (2000) framework. Much of this work has focused on the question of whether or not people 'backtrack'. For example, consider a causal chain structure $A \rightarrow B \rightarrow C$ in which none of the events happened. A backtracking counterfactual question asks whether A would have happened if one had intervened to make B happen. According to Pearl's framework the answer is 'no' (i.e. $p(a_b|a', b', c') = 0$). Because counterfactuals are construed as interventions that break any incoming links into the intervened-on variable, only the values of variables that are downstream from the intervention may change. There is no backtracking in Pearl's framework (i.e. changing the values of variables upstream from the intervention). However, several studies have shown that people sometimes do backtrack (Dehghani, Iliev, & Kaufmann, 2012; Gerstenberg, Bechlivanidis, & Lagnado, 2013; Hiddleston, 2005; Lucas & Kemp, 2015; Rips, 2010; Rips & Edwards, 2013). There is also a rich literature on the development of counterfactual reasoning, which I will say more about in the General Discussion.

Only a few studies have looked directly into whether people distinguish between the second and third level of the causal hierarchy. Meder, Hagmayer, and Waldmann (2009) studied what causal inferences participants draw based on evidence from observations (level I), evidence from interventions (level II), or evidence from counterfactuals (level III). The experiment was designed such that, normatively, different inferences were licensed for each kind of evidence. While the results showed that participants differentiated between observational and interventional evidence, they didn't distinguish counterfactual from interventional evidence.

Most relevant to the question of how hypothetical and counterfactual judgments relate to causal judgments is a recent paper by Skovgaard-Olsen, Stephan, and Waldmann (2021). Across a series of six experiments, the authors show that people differentiate between indicative conditionals ("if x happens then y happens") and counterfactual conditionals ("if x had happened then y would have happened"). For example, consider a situation in which A is a common cause of both B and C, $B \leftarrow A \rightarrow C$. In this case, the indicative conditional "if b happens, then c happens" is true (i.e. $p(b|c)$ is high). But the counterfactual conditional "if b had happened, then c would have happened" is false (i.e. $p(c_b|b', c')$ is low).

Considering a counterfactual intervention that had changed B wouldn't have affected C.

In their Experiment 5, Skovgaard-Olsen et al. (2021) used such a common-cause structure to test whether participants' judgments about indicative conditionals and counterfactual conditionals come apart and, if so, which type of conditional better aligns with causal judgments. Participants were either asked about the relationship between A and B (predictive), B and A (diagnostic), or B and C (spurious). Each participant judged the probability of an indicative conditional being true (e.g. "if a then b" in the predictive condition), a counterfactual being true (e.g. "if a' then b'", phrased as "if a had not happened, then b would not have happened."), or a causal statement being true (e.g. "a caused b"). The results showed that participants responded differently to the different question types. Whereas their responses for indicative conditionals were essentially the same in each of the three conditions (predictive, diagnostic, spurious), for counterfactual conditionals and causal statements their answers differed between the conditions. For example, they said that A caused B in the predictive condition, but that A didn't cause B in the diagnostic condition. Importantly, participants' counterfactual and causal judgments were closely aligned with one another, whereas participants' judgments about the indicative conditionals didn't match their causal judgments.

The role of counterfactuals in theories of causal judgment. Much prior research has argued that counterfactuals and causal judgments are intimately linked (Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Kahneman & Tversky, 1982). Here, by causal judgments, I mean judgments about what caused what to happen in a particular situation, such as whether the sprinkler caused the grass to be wet. As we have seen, counterfactuals also form the basis for recent approaches in computer science that formally model causal judgments (Halpern, 2016; Pearl, 2000). In these approaches, causal knowledge is expressed in the form of causal Bayes nets or structural equations that capture the causal dependence between the variables in the model. Counterfactuals are construed as interventions that set a variable to a desired value (see also Hitchcock, 2001; Woodward, 2003, 2021; Yablo, 2002). By considering such counterfactual interventions, the formalism yields verdicts about which variables actually caused some outcome of interest (Halpern & Pearl, 2005). Intuitively, it's those variables that were pivotal for the outcome which caused it to come about (see Chockler & Halpern, 2004; Gerstenberg et al., 2018; Lagnado, Gerstenberg, & Zultan, 2013; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021, for work showing how this idea of being pivotal is important for judgments of responsibility as well).

While a simple counterfactual test fails in situations of causal overdetermination (where two or more individually sufficient causes brought about an outcome), more sophisticated tests have been developed to deal with such situations (see Halpern, 2016; Halpern & Pearl, 2005, for details). These tests consider not only whether a variable was pivotal in the actual situation, but also whether it would have been pivotal in other possible situations that could have arisen.[2]

An alternative class of approaches – process theories of causation – explains causal judgments merely in terms of what actually happened and without relying on counterfactu-

[2]The challenge is then to impose restrictions onto what possible situations may be considered in such a way that the verdicts of this formalism agree with people's intuitions about which events caused the outcome (see Halpern, 2016, for details).

als (Talmy, 1988; Wolff, 2007). Empirical work has shown that people's causal judgments are indeed sensitive to the way in which the outcome came about, and not just to mere counterfactual dependence (Lombrozo, 2010; Mandel, 2003; Shultz, 1982; Walsh & Sloman, 2011; Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010).

Inspired by both counterfactual theories and process theories of causation, Gerstenberg et al. (2021) developed the counterfactual simulation model (CSM) of causal judgment for physical events. The CSM predicts that people's causal judgments are a function of their subjective degree of belief that the candidate cause made a difference to whether or not the outcome of interest happened. The dashed paths in Figure 1 show how ball B would have moved if ball A hadn't been present in the scene. An observer doesn't have direct access to what would have happened. Instead, they need to use their intuitive understanding of the domain to simulate the counterfactual. Gerstenberg et al. showed that the CSM accurately captured people's quantitative causal judgments. The more certain participants were that ball A's presence made a difference to whether or not the outcome happened, the more they agreed that ball A caused ball B to go through the gate. So, for example, participants gave high causal ratings for clips like the one in Figure 1a, and low causal ratings for clips like the one in Figure 1b. Participants gave intermediate judgments whenever it was unclear whether ball B would have gone through the gate if ball A hadn't been present in the scene (i.e. when ball B was initially headed to one of the edges of the gate).

Direct evidence for spontaneous counterfactual simulation? Gerstenberg et al. (2021) show that the CSM captures people's causal judgments to a high degree of quantitative accuracy. However, they don't show directly that people engage in *counterfactual* simulation when making causal judgments. By tracking participants' eye-movements, Gerstenberg et al. (2017) demonstrated that participants spontaneously tried to assess where ball B would go if ball A wasn't present in the scene. When asked to make causal judgments, participants didn't just focus their attention on what actually happened. Instead, their eyes saccaded to where ball B would go if ball A wasn't present. Saccades are fast eye-movements from one place to another that exceed a certain velocity threshold. These eye-movements were more frequent in situations in which the counterfactual outcome was less clear (i.e. situations in which ball B was headed toward one of the edges of the gate), suggesting they may serve the purpose of reducing uncertainty about the counterfactual outcome. In contrast, when participants were asked to make a judgment about the actual outcome (i.e. how closely ball B went through the gate, or missed the gate), they tended to focus on what actually happened and only rarely saccaded to where ball B would have gone. Looks to where ball B would have gone were recruited specifically in service of making causal judgments.

Figure 4 shows the endpoints of participants' saccades for one of the clips from the experiment, separated by the experimental condition. The conditions only differed in terms of what questions participants were asked to answer about the clip. Gerstenberg et al. termed those looks "counterfactual saccades" for which the endpoint of the saccade was close to the path that ball B would have taken if ball A had been absent. The results showed that participants produced more counterfactual saccades in the counterfactual condition (where they were asked to say whether ball B would have gone into the gate if ball A hadn't been there), and in the causal condition, compared with the outcome condition.

While Gerstenberg et al. termed these looks "counterfactual saccades" it's important

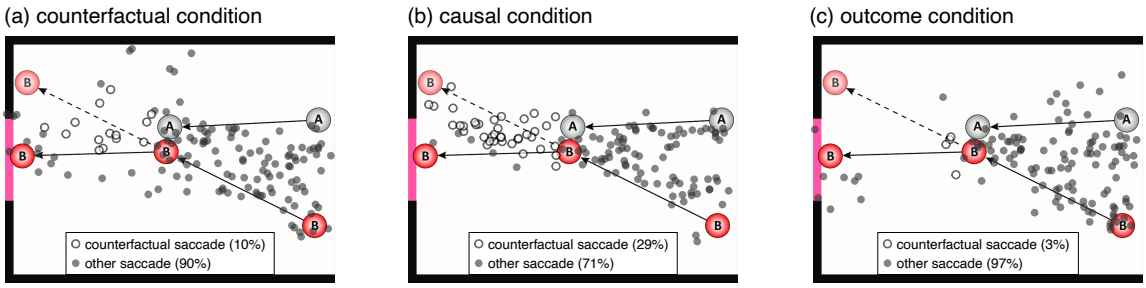

Figure 4. Saccade plots for one of the trials from Gerstenberg et al. (2017) in the counterfactual, causal, and outcome condition. Each point in the plot shows an endpoint of a saccade (a fast eye-movement from one position to another). Saccade endpoints that were both close to the counterfactual path that ball B would have taken if ball A hadn't been present in the scene, and far enough to the left of where the collision happened, were classified as "counterfactual saccade" (white points). The rest were classifed as "other saccade" (black points). The time window for this analysis was constrained to range from after the two balls entered the scene to before they collided with one another. This was done because after the two balls collide, ball A travels on a similar path to the one that ball B would have taken if ball A hadn't been present in the scene. By restricting the time window to before the collision, one can be sure that the "Counterfactual saccades" are anticipatory saccades to where ball B would go, rather than saccades to where ball A currently is.

to note that these looks actually happened *before* the two balls collided. So, in some sense, these looks were "hypothetical saccades" to where ball B would go if ball A were removed from the scene. The CSM postulates that people make causal judgments by comparing what actually happened with what they believe *would have happened* in the relevant counterfactual situation, such as in the situation in which ball A had been removed from the scene. Another possibility, however, is that people compute the probability of a future hypothetical outcome instead, and then compare what actually happened to that hypothetical outcome.

As discussed earlier, Skovgaard-Olsen et al. (2021) show that people are sensitive to the difference between inferences on level I and level III on Pearl's (2000) causal hierarchy, and that level III inferences are more closely aligned with causal judgments than level I inferences. The work presented here is a natural follow up. I will show that level III inferences are critical for capturing causal judgments about dynamic physical interactions, and that level II inferences don't suffice. I will present a computational model that implements hypothetical and counterfactual inference as mental simulations in a physical setting, and then test in two experiments which kind of mental simulation better explains people's causal judgments.

Simulation Model

In the experiments below I ask different groups of participants to make three different kinds of judgments: hypothetical judgments about what would happen in the future, counterfactual judgments about what would have happened if things had been different, or causal judgments. Figure 5a shows an example of the kinds of video clips that participants

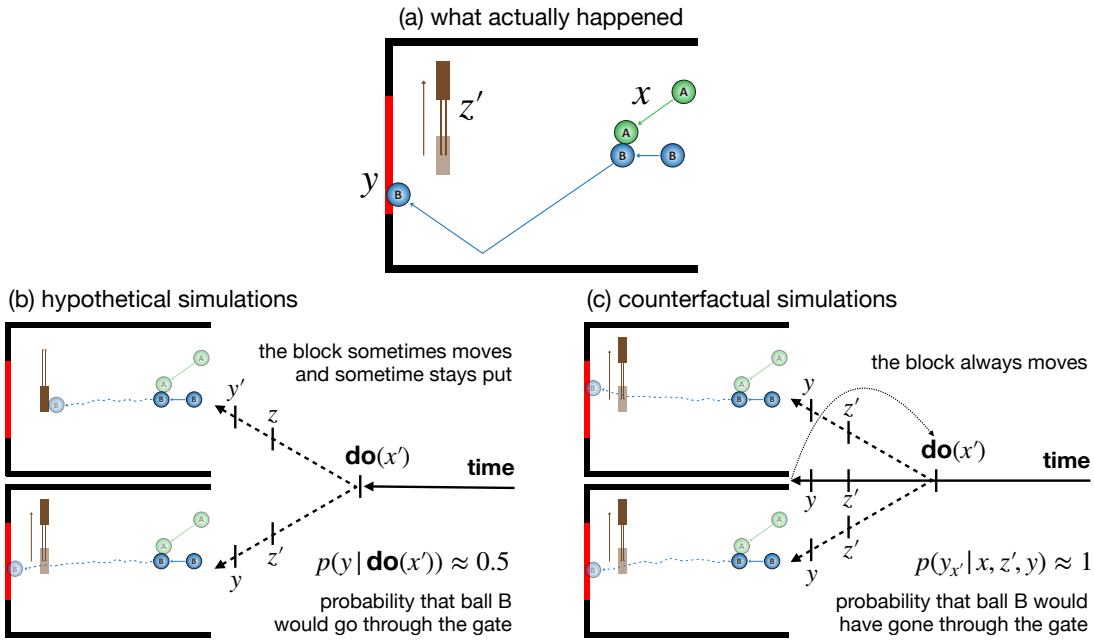

Figure 5. Diagrams illustrating what actually happened as well as how hypothetical simulations and counterfactual simulations are generated. a) In the actual situation, ball A was present in the scene (x), the block didn't end up in front of the gate (z'), and ball B went through the gate (y). b) To compute the hypothetical probability of whether ball B would miss the gate if ball A weren't present in the scene $p(y'|\text{do}(x'))$, the model removes ball A from the scene, and then simulates what would happen. The dashed path illustrates ball B's movement in the simulation. At each moment in the simulation, a small degree of noise is added to ball B's trajectory to capture the fact that participants have some degree of uncertainty about exactly how ball B would move if ball A weren't there. In some of the simulations the block stays put (see top example) while in others the block moves (bottom example). c) To compute the counterfactual probability of whether ball B would have missed the gate if ball A hadn't been present in the scene $p(y'_{x'}|x, y)$, the model first takes into account everything that actually happened which includes whether or not the block moved. It then goes back in time (indicated by the dotted arrow in the diagram) to replay the clip with ball A removed. Based on the outcome of many such simulations, the *hypothetical probability* with which ball B would go through the gate if ball A weren't present is around 50% because the block moves with 50% probability. The *counterfactual probability* with which ball B would have gone through the gate if ball A hadn't been present in the scene is close to 100% because the block always moves in each counterfactual simulation just like it did in the actual situation.

saw in the experiments. I will represent this clip with three variables. X denotes whether ball A was present in the scene (x) or not (x'). Y denotes whether ball B went through the gate (y) or didn't go through (y'), and Z denotes whether the final position of the block was in front of the gate (z) or out of the way (z').

In the hypothetical condition, the model computes the probability of whether ball B

would go through the gate if ball A weren't present in the scene, $p(y|\text{do}(x'))$. The model computes this probability by running a number of hypothetical simulations. The model first conditions on what it actually observed. That is, it considers the balls' initial trajectories and the initial state of the block. The model then removes ball A from the scene and simulates what would happen in its absence. Figure 5b shows two runs of the simulation model in the hypothetical condition. In the top one, ball B didn't go into gate because the block stayed put. In the bottom one, ball B ended up going into gate because the block moved out of the way.

In the counterfactual condition, the model computes the probability of whether ball B would have gone through the gate if ball A hadn't been present in the scene, $p(y_{x'}|x, y)$. Here, instead of only conditioning up to the point at which the two balls collided in the actual situation, the model takes into account the full clip until the end. So it observes whether or not ball B went into the gate, and also whether or not (and when) the block moved. When the model now simulates what would have happened if ball A hadn't been present in the scene, it still has uncertainty about ball B's movement, as well as about the exact time at which the block would have moved. However, importantly, it doesn't have any uncertainty about whether or not the block moved. Figure 5c shows two runs of the simulation model in the counterfactual condition. In both situations, ball B ends up going through the gate, as the block always moves out of the way (just like it did in the actual situation).

The simulation model has two sources of uncertainty. First, the model has uncertainty about ball B's movement. To capture this uncertainty, the model introduces noise to ball B's movement at each time step in the physics simulation by applying a small perturbation to the direction of the ball's velocity vector. At each time step in the simulation, the direction of ball B's velocity is randomly perturbed. The perturbation is drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma_{\text{ball}})$, where σ_{ball} affects how strong the random perturbations are.

Second, the model has uncertainty about whether, and if so, when the block moves. In the experiment, I tell participants that the block sometimes moves and sometimes stays put. Across the clips that participants see in the experiment, the block moves half of the time. So, in the model, I set the probability that the block will move to $p = 0.5$. If the block moves, then the model is uncertain about exactly when the block will start moving. I model this uncertainty by adding noise to the actual time point in the physical simulation at which the block moves. This noise is drawn from a Gaussian distribution with $\mathcal{N}(0, \sigma_{\text{block}})$, where σ_{block} determines the degree of uncertainty in when the block would move. The time point at which the block starts moving is constrained to lie between the time at which the two balls collide, and the time point at which the clip ends. This constraint makes it so that it's not possible in a hypothetical simulation for the block to start moving before the time point at which the balls collided in the actual situation. Remember that the model conditions on what happened up to this point at which the block is still in its initial position.

To compute the hypothetical probability $p(y|\text{do}(x'))$ and the counterfactual probability $p(y_{x'}|x, y)$, the model generates a large number of simulations for each clip and then simply computes the proportion of simulations in which ball B ended up going through the gate. The key difference between hypothetical and counterfactual simulations is that for counterfactual simulations, the model conditions on what actually happened all the way until the end of the clip. In contrast, for hypothetical simulations, the model only condi-

tions on what happened until before the two balls collided. This means that the model has more uncertainty about the hypothetical outcome (because the block may or may not move) compared to the counterfactual outcome. For example, for the clip shown in Figure 5a, the hypothetical probability is close to 50% (ball B only goes through if the block goes out of the way), whereas the counterfactual probability is close to 100% (the block moves in each of the simulations but it might sometimes not move out of the way in time).

The simulation models then uses the hypothetical or counterfactual probabilities to compute the probability that x caused y. According to the *hypothetical simulation model*, the probability that x caused y is given by

$$p(x \rightarrow y) = p(y'|\text{do}(x')). \tag{1}$$

The more likely ball B would miss the gate (y') if ball A weren't there ($\text{do}(x')$), the more likely ball A caused ball B to go into the gate.

According to the *counterfactual simulation model*, the probability that x caused y is given by

$$p(x \rightarrow y) = p(y'_{x'}|x, y). \tag{2}$$

The more likely ball B would have missed the gate if ball A hadn't been there ($y'_{x'}$), when it fact ball A was present (x) and ball B went into the gate (y), the more likely ball A caused ball B to go into the gate.[3]

Experiment 1 tests whether the simulation model accurately captures participants' hypothetical and counterfactual judgments. Experiment 2 then tests whether participants' causal judgments are better explained by hypothetical simulations (Equation 1), or by counterfactual simulations (Equation 2).

Experiment 1: Hypothetical vs. counterfactual simulations

In this experiment, I wanted to see how people make hypothetical and counterfactual judgments, and whether their judgments would be accurately captured by the simulation model.

Methods

All of the materials including the data, experiment code, and analysis scripts are available here: `https://github.com/cicl-stanford/counterfactual_hypothetical/`

Participants. 110 participants (*age*: M = 35, SD = 10; *gender*: 35 female, 72 male, 1 non-binary, 2 preferred not to say; *race*: 87 White, 11 Black, 10 Asian, 1 Native American, 1 preferred not to say; *ethnicity*: 13 Hispanic, 96 not Hispanic, 1 preferred not to say) were recruited via Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016). Only participants based in the US with an approval rating of 95% or higher were able to participate (Mason & Suri, 2012).

[3] To answer the question of whether ball A prevented ball B from going through the gate, the model computes the probability that ball A caused ball B to miss. For the *hypothetical simulation model*, this probability is given by $p(x \rightarrow y') = p(y|\text{do}(x'))$, and for the *counterfactual simulation model*, it is given by $p(x \rightarrow y') = p(y_{x'}|x, y')$.

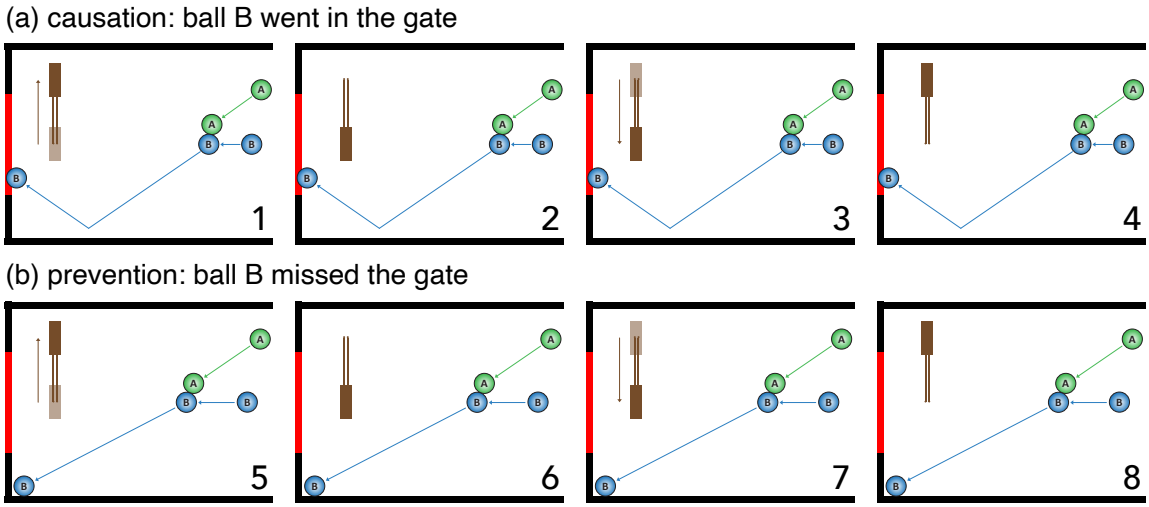

Figure 6. Diagrams of the eight test clips that participants saw. In each clip, both ball A and ball B are initially out of sight, enter the scene from the right, and collide with one another. The diagrams show the full trajectory of ball B, and the trajectory of ball A up until the collision (what trajectory ball A took after the collision is not shown). (a) In clips 1–4, ball B goes through the gate on the left. (b) In clips 5–8, ball B misses the gate. In half of the clips, the brown block in front of the gate moves its position, whereas in the other half it stays put. For example, in clip 1 the block is initially in front of the gate, and then moves up. In this clip, ball B would have gone through the gate even if ball A hadn't been present in the scene (because the block moved out of the way before ball B would have gotten there). In the hypothetical condition, the clip paused shortly before the two balls collided. The block was still at its initial position at this point in time. In the counterfactual and causal condition, the clip played until the end.

Design and Procedure. The experiment had two conditions that differed in whether participants were asked to answer a hypothetical question about what would happen if ball A was removed, or a counterfactual question about what would have happened if ball A had been removed.

The instructions in both conditions were largely identical. Participants were told that their task would be to make judgments about video clips, and they viewed two diagrams similar to those in Figure 6 illustrating what the clips will look like. Participants learned that in each clip, two balls, ball A and ball B, enter the scene from the right and collide with one another. In some of the clips, ball B ends up going through the red gate on the left, and in some of the clips ball B misses the gate. They were also told that there is a brown block on a track that may or may not move. They weren't provided with any specific information about how likely it was that the block would move and, if so, at what time. In fact the block moved in half of the clips and stayed put in the other half. For the clips in which it moved, it moved at the same point in time.

Participants were asked a set of multiple choice comprehension check questions. One of the questions made sure that participants had paid attention to the fact that the block sometimes slides along the track, and that it sometimes stays put. If any of the com-

prehension check questions were answered incorrectly, participants were redirected to read the instructions again. Only participants who answered all of the comprehension check questions correctly were able to proceed to the test phase.

Participants first watched two practice clips to familiarize themselves with the setting and the task. In both of the clips, the two balls collided with one another. In one of the clips, ball B didn't go through the gate and the block moved. In the other clip, ball B went through the gate and the block didn't move. After the practice clips, the eight test clips shown in Figure 6 were presented in randomized order. In all of the clips, balls A and B enter the scene from the right and collide with one another. In clips 1-4, ball B goes through the gate (top row). In clips 5-8 ball B doesn't go through the gate (bottom row). As the figure shows, ball B's full trajectory and ball A's trajectory up until the two balls collided were identical within each of two sets of clips. What differed between the clips was the initial position of the block, and whether or not it moved. For example, in clip 1, the block was initially in front of the gate but then moved out of the way. In clip 2, the block didn't move and stayed in front of the gate the whole time.

In the *hypothetical condition* ($N = 50$), each clip paused shortly before the two balls collided with one another. Participants were allowed to replay the clip as many times as they liked. They were then asked to what extent they agreed with the statement "Ball B would go through the gate if ball A wasn't there" and indicated their answer on a sliding scale with the endpoints being labeled "not at all" (0) and "very much" (100). After having provided their judgment on the slider, participants viewed the full clip of what actually happened until the end. The reason I wanted participants to see the whole clip until the end was so that they could see that the block sometimes moved, and sometimes didn't move. Note that they didn't get to see a clip of what would have happened if ball A hadn't been there, so they didn't get direct feedback about whether or not their judgment was correct.

In the *counterfactual condition* ($N = 60$), each clip played until the end. Here, too, participants were able to replay the clips as many times as they liked. They were then asked to what extent they agreed with the statement "Ball B would have gone through the gate if ball A hadn't been there" and indicated their answer on a sliding scale with the endpoints being labeled "not at all" (0) and "very much" (100). So the key differences between the hypothetical and the counterfactual condition were how the question was phrased, and at what time point in the clip the question was asked – shortly before the collision in the hypothetical condition, or at the end of the clip in the counterfactual condition.

After the test phase participants provided demographic information. They were also asked what factors influenced how they made their judgment and responded using a free text form. On average, it took participants 7.9 minutes (SD = 3.6) to complete the experiment.

Results

Figure 7 shows participants' ratings in the hypothetical condition and in the counterfactual condition for the eight test clips, together with the model predictions. I'll discuss the results from the hypothetical condition and the counterfactual condition in turn before comparing participants' ratings with the model predictions.

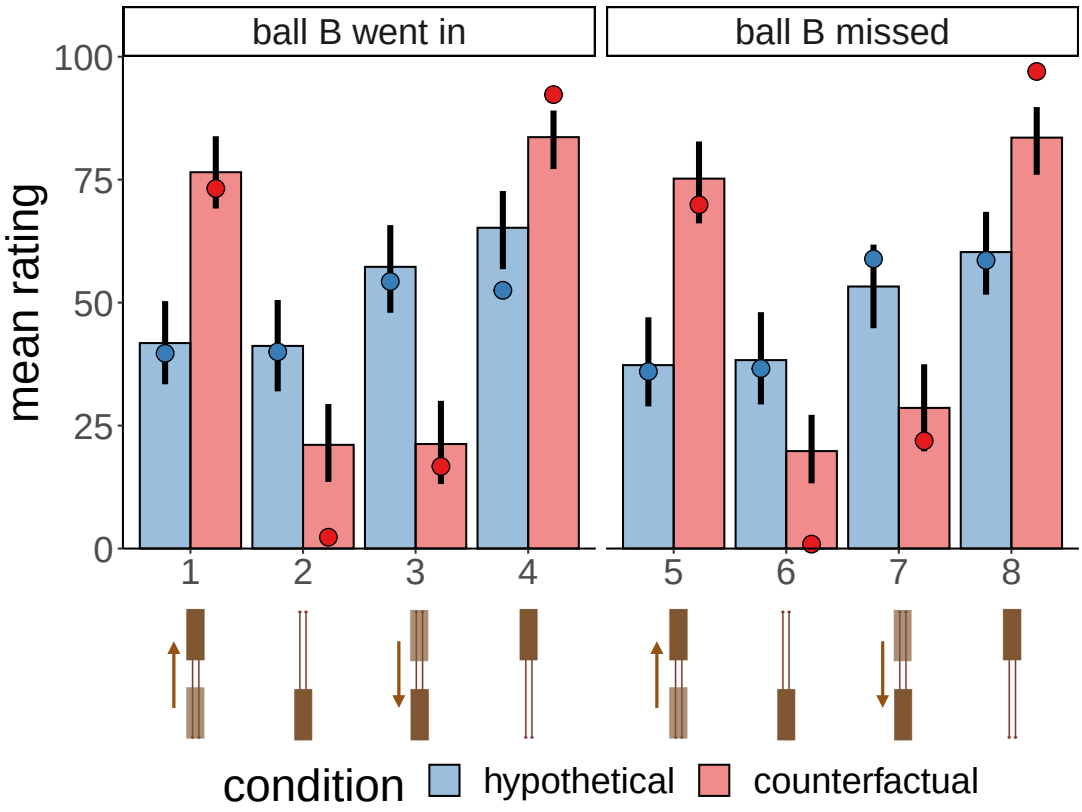

Figure 7. Mean ratings (bars) in the hypothetical (blue) and counterfactual (red) condition for situations in which ball B went in (left) or ball B missed (right), together with the model predictions (circles). The images on the x-axis illustrate the initial and final position of the block. The block is in front of the gate if it's at the bottom. For example, in clip 1, the block was initially in front of the gate but then moved out of the way *Note*: Error bars are 95% bootstrapped confidence intervals.

Hypothetical condition

Figure 7 shows that participants' mean judgments in the hypothetical condition tended to be close to the midpoint of the scale, and that they were affected only by the initial position of the block (see Table 2, block initial).[4] For example, judgments of whether ball B would go through the gate if ball A weren't there were lower in clips 1 and 2 where the block was initially in the way than they were in clips 3 and 4 for which the block was initially out of the way. Recall that in the hypothetical condition, participants only viewed the clip up until shortly before the collision at which point the clip paused. Thereby, it is no surprise that the final position of the block didn't affect their judgments (as they couldn't have known at the time of making the judgment what the final position of the block would be). Whether ball B ended up going into the gate, or whether it missed the gate also didn't affect participants' hypothetical judgments. Again, at the time of judgment participants

[4] I will refer to a factor as having influenced participants' judgments when the 95% credible interval of the posterior distribution for that factor excludes 0.

didn't know what the outcome would be.

Counterfactual condition

Participants' counterfactual judgments were most strongly affected by the final position of the block (see Table 2, block final). For example, they agreed that ball B would have gone into the gate if ball A hadn't been present when the block moved out of the way (clip 1) or had stayed out of the way (clip 4), but disagreed when the block didn't move out of the way (clip 2) or moved into the way (clip 3). The same pattern of judgments holds for those clips in which ball B missed the gate.

While the final position of the block was most important for participants' counterfactual judgments, there was also a small effect of the initial position of the block (see Table 2, block initial). Participants' counterfactual judgments tended to be higher when the block had stayed out of the way the whole time (clips 4 and 8) compared to when the block moved out of the way (clips 1 and 5). Similarly, their judgments tended to be lower when the block was in front of the gate the whole time (clips 2 and 6) compared to when it moved into the way (clips 3 and 7).

Model comparison

As Figure 7 shows, the simulation model closely tracks participants' hypothetical and counterfactual judgments (Pearson correlation: $r = .97$, Spearman correlation: $r_s = .79$, Root mean squared error: RMSE = 12.50). The simulation model has two free parameters: one that captures participants' uncertainty in how exactly ball B would have moved if ball A hadn't been present in the scene σ_{ball}, and one that captures participants' uncertainty about the moment in time in which the block would have started to move σ_{block}. I fitted these two parameters in the model to participants' judgments by minimizing the sum of squared errors between model predictions and participants' mean hypothetical and counterfactual judgments (see Appendix for more details on how the parameters were fitted). This analysis reveals that uncertainty about when the block will move is more important

Table 2
Posterior means and 95% highest density intervals for each fixed effect in the Bayesian mixed-effects regression model. I fitted the model separately for participants' hypothetical and counterfactual judgments. The results shows that participants' hypothetical judgments are most strongly influenced by the initial position of the block, and that counterfactual judgments are mostly strongly influenced by the final position of the block. Note*: I used sum contrasts for the predictor variables with no/yes for the block variables, and miss/hit for the outcome variable.*

model specification: `judgment ~ 1 + block_initial + block_final + outcome + (1 | participant)`

name	intercept	block initial	block final	outcome
hypothetical	49.26 [44.95, 53.61]	9.67 [6.64, 12.66]	1.84 [−1.08, 4.94]	−2.08 [−5.3, 1.03]
counterfactual	51.2 [48.02, 54.46]	3.05 [0.48, 5.75]	28.51 [25.91, 31.11]	0.59 [−2.03, 3.29]

for capturing participants' judgments than the uncertainty that is associated with the ball's movement trajectory. As ball B travels on a straight horizontal path towards the middle of the gate, extrapolating how it would have moved if ball A hadn't been present is fairly straightforward. In contrast, remembering at what moment in time the block moved and whether it would have moved early enough so as to get out of the way (or into the way) is more difficult to assess.

It's worth noting that the simulation model captures the effect that the initial position of the block has on participants' counterfactual judgments. For example, participants' counterfactual judgments are slightly higher in clips 4 and 8 than in clips 1 and 5. When the block is initially in the way, there is some chance that it's not going to move out of the way in time. So the chances of ball B being blocked is a little greater in clips 4 and 8 than in clips 1 and 5. On the other hand, the model predicts a larger difference between counterfactual judgments for clips 2 and 6 versus clips 3 and 7 than what was observed. When the block stays put in front of the gate, there is only a very small chance that ball B would have gone through the gate if ball A hadn't been present in the scene. While the model predicts a very low rating in this case, participants' judgments were a little higher.

Discussion

The results of Experiment 1 show that hypothetical and counterfactual judgments come apart in this paradigm. In the hypothetical condition, the clips paused shortly before the collision between the balls and participants judged whether ball B would go through the gate if ball A wasn't present in the scene. Here, participants' judgments were only affected by where the block was positioned at the time when video clip paused. Participants' hypothetical judgments that ball B would go through were a little higher when the block was initially out of the way than when it was in the way. I had told participants in the instructions that the block sometimes moves. While in fact the probability that the block moved was 50% across the ten clips that participants saw, they did not know this. So it makes sense that they would assume that ball B would be less likely to go into the gate when the block was initially in the way than when it was out of the way.[5]

In the counterfactual condition, the clips played until the end so participants were able to see whether or not the block moved in the actual situation. Here, participants' judgments were most strongly affected by the final position of the block. They agreed that ball B would have gone through the gate if ball A hadn't been present when the final position of the block was out of the way, and disagreed when the block ended up in front of the gate. Their judgments were also somewhat affected by the initial position of the block. Counterfactual judgments were higher when the block was out of the way the hole time, and lower when the block was in front of the gate the whole time (compared to when it moved).

The reason that hypothetical and counterfactual judgments come apart in this paradigm is that the truth of the hypothetical (or counterfactual) depends on a factor that is independent of the causal event of interest (ball A's presence). Whereas in the counterfactual condition, participants get to see whether or not the block moved, partici-

[5]Participants were able to learn how likely the block is to move over the course of the experiment as they got to see the full video clip after having made their hypothetical prediction.

pants in the hypothetical condition don't know. This makes it such that the hypothetical of what would happen if ball A wasn't present is different from the counterfactual of what would have happened if ball A hadn't been present. Participants in the hypothetical condition were more uncertain about what would happen than participants in the counterfactual condition were about what would have happened.

The simulation model accurately captured participants' hypothetical and counterfactual judgments. The model assumes that observers may be uncertain about how exactly ball B would have moved in the absence of ball A, and at what point in time the block would have started to move. These two sources of uncertainty are sufficient to explain the overall pattern of responses.

Experiment 2: Causal judgments

Experiment 1 established that hypothetical and counterfactual judgments come apart in this paradigm, and that the simulation model captures both kinds of judgments. In Experiment 2, I asked participants to make a causal judgment about what happened in each clip. Specifically, whether ball A caused ball B to go through the gate (when it went through), and whether ball A prevented ball B from going through the gate (when it missed). Will participants' causal judgments be better explained assuming that they compare what actually happened to the simulation of a future hypothetical, or to the simulation of a counterfactual?

Methods

Participants. 67 participants (*age*: M = 34, SD = 10; *gender*: 20 female, 46 male, 1 non-binary; *ethnicity*: 9 Hispanic, 57 not Hispanic, 1 preferred not to say) were recruited via Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016). Only participants based in the US with an approval rating of 95% or higher were able to participate.

Design and procedure. The procedure was largely identical to that of the counterfactual condition in Experiment 1. The only thing that differed was what questions participants were asked. Participants were asked to what extent they agree with the statement "Ball A has caused ball B to go through the gate" if ball B went through the gate, or "Ball A has prevented ball B from going through the gate" if ball B didn't go through the gate. Participants indicated their answer on a sliding scale with the endpoints labeled "not at all" (0) and "very much" (100). Just like in the counterfactual condition of Experiment 1, participants watched the clip until the end before providing their causal judgment.

Results

Figure 8 shows participants' mean causal judgments together with the model predictions based on participants' hypothetical and counterfactual judgments from Experiment 1. Participants' causal judgments are explained by an interaction between the final position of the block and the outcome (see Table 3). For situations in which ball B went in the gate, participants' causal judgments were high when the final position of the block was in front of the gate (clips 2 and 3), and low when the block wasn't in front of the gate (clips 1 and 4). Conversely, when ball B missed the gate, participants' judgments were high when the final position of the block was *not* in front of the gate (clips 5 and 8), and low when the

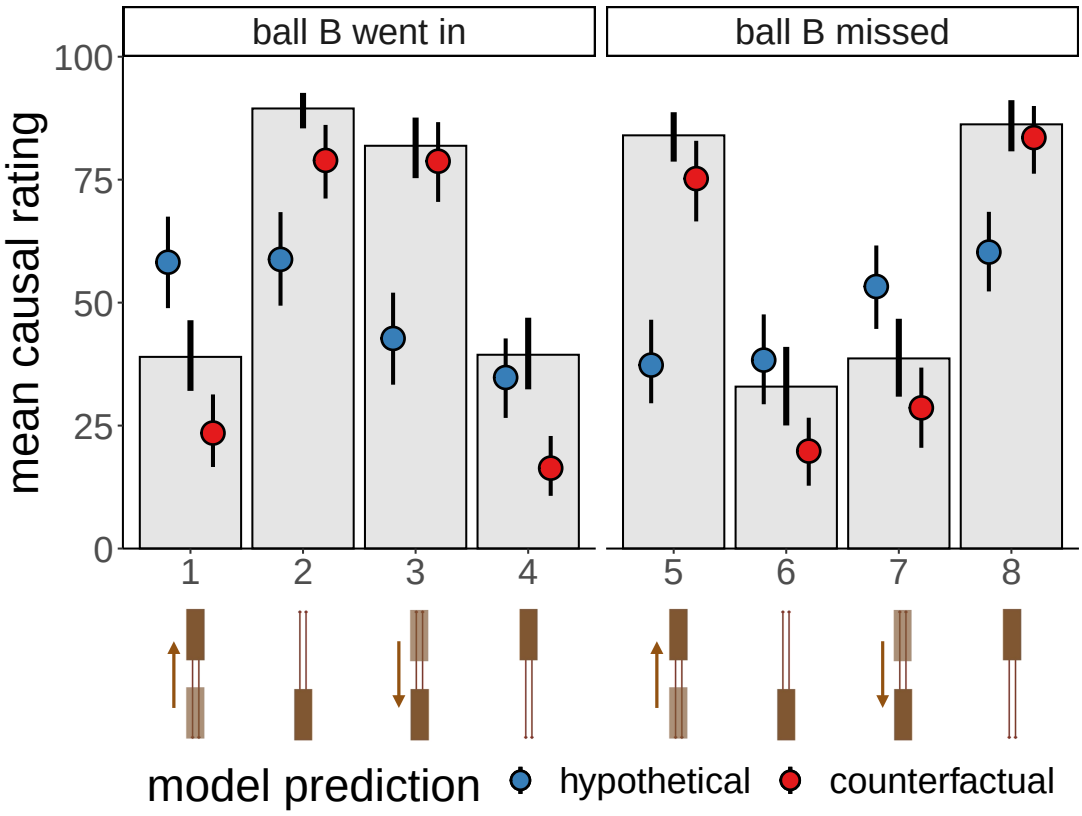

Figure 8. Mean causal ratings (gray bars) as a function of whether ball B went in (left) or ball B missed the gate (right). Model predictions based on participants' judgments in the hypothetical condition (blue) and the counterfactual condition (red) are shown as circles. The images on the x-axis illustrate the initial and final position of the block. The block is in front of the gate if it's at the bottom. For example, in clip 1, the block was initially in front of the gate but then moved out of the way. Notice that the model predictions for when ball B went in are flipped compared to those shown in Figure 7. This is because in Experiment 1, we asked participants whether ball B would go in (hypothetical), or would have gone in (counterfactual) without ball A. But here, the model uses the probability that ball B would *not* go in (hypothetical), or would *not* have gone in (counterfactual) without ball A (see Equation 1 and 2 and as well as Footnote 6). *Note*: Error bars are 95% bootstrapped confidence intervals.

block was in front of the gate (clips 6 and 7). The initial position of the block had little to no effect on participants' causal judgments. What mattered was whether ball B ended up in the gate, and whether the block would have blocked ball B if ball A hadn't been present in the scene.

As Figure 8 also shows, participants' causal judgments in Experiment 2 lined up nicely with participants' counterfactual judgments in Experiment 1. Participants' mean counterfactual and causal judgments were highly correlated with one another ($r = .99, r_s = .79, \text{RMSE} = 12.50$). In contrast, the correlation between mean hypothetical and causal

judgments was much lower ($r = .24, r_s = .40, \text{RMSE} = 27.31$).[6]

Discussion

The results of Experiment 2 demonstrate that causal judgments are best explained by counterfactuals and not by hypotheticals. Participants' causal judgments in Experiment 2 closely aligned with participants' counterfactual judgments in Experiment 1, whereas the correlation between causal judgments and hypothetical judgments was much lower. For example, when judging whether ball A caused ball B to go through the gate, it matters what would have happened if ball A hadn't been present in the scene. If ball B would have gone through the gate even if ball A hadn't been present in the scene (because there was no block in the way), then people tended to disagree with the statement that ball A caused ball B to go through the gate. However, when ball B would have been blocked had ball A not been presented in the scene, then participants' tended to agree that ball A caused ball B to go through the gate. The same holds for participants' judgments about whether ball A prevented ball B from going through the gate. They agreed when ball B would have gone in in the absence of ball A (i.e. when it wouldn't have been blocked), but disagreed when ball B would have missed even if ball A had been removed (because the gate was blocked).

General discussion

This paper asked the question of whether counterfactuals are necessary for explaining causal judgments, or whether comparing what actually happened with a future hypothetical simulation suffices. The answer is clear: counterfactuals are necessary. People make causal judgments about particular events by comparing what actually happened with what would have happened in a counterfactual situation (Gerstenberg et al., 2021).

Prior work had tested the idea that causal judgments are intimately linked to counterfactual simulations and found a close fit between causal and counterfactual judgments

Table 3
Posterior means and 95% highest density intervals for each fixed effect in the Bayesian mixed-effects regression model. The results show that the interaction between the final position of the block and the outcome ("block final : outcome") predicts most of the variance in participants' causal judgments. Note*: I used sum contrasts for the predictor variables with no/yes for the block variables, and miss/hit for the outcome variable.*

model specification: `judgment ~ 1 + (block_initial + block_final) * outcome + (1 | participant)`

intercept	block initial	block final	outcome	block initial : outcome	block final : outcome
61.46 [58.1, 64.42]	0.11 [−2.31, 2.22]	0.71 [−1.58, 2.78]	−1.02 [−3.22, 1.07]	1.91 [−0.3, 4.25]	23.97 [21.72, 26.13]

[6] Before correlating the judgments with one another, I subtracted participants' hypothetical and counterfactual judgments from 100 for the situations in which ball B went through the gate. Participants in Experiment 1 were asked to judge whether ball B would (or would have gone) through the gate if ball A weren't (or hadn't been) there. To map these onto participants' causal judgments in Experiment 2 when ball B went through the gate, we need participants' hypothetical and counterfactual judgments that ball B would *not* go (or would *not* have gone) through the gate (see Equations 1 and 2).

(Gerstenberg et al., 2017). Participants' judgments that ball A caused ball B to go through the gate were higher the more certain they were that ball B wouldn't have gone through if ball A hadn't been there. However, the results of these studies are consistent with the idea that people compare what actually happened with the outcome of a hypothetical future situation. In order to tease hypothetical and counterfactual probabilities apart, situations are required in which multiple factors influence the outcome, and where the observer has some uncertainty about at least one of these factors. I generated video clips in which hypotheticals and counterfactuals come apart by introducing a block in front of the gate that sometimes moves and sometimes stays put.

Experiment 1 asked one group of participants to make hypothetical judgments about whether ball B would go through the gate if ball A weren't there, and another groups of participants to make counterfactual judgments about whether ball B would have gone through the gate if ball A hadn't been there. As predicted, participants' hypothetical and counterfactual judgments came apart in these clips and they were well captured by a computational model that computes hypothetical and counterfactual probabilities by running noisy physical simulations that incorporate people's uncertainty about what would or would have happened. Experiment 2 then asked participants to make causal judgments. The results showed that participants' causal judgments in Experiment 2 were closely in line with participants' counterfactual judgments from Experiment 1. In contrast, the hypothetical judgments from Experiment 1 didn't capture participants' causal judgments as well.

In the remainder, I will discuss what implications these results have for theories of causality, and for research into the development of counterfactual reasoning. I conclude by pointing out some limitations of the work presented here that motivate possible (hypothetical) future directions.

Implications for theories of causality

The term "counterfactual" is often used quite broadly to refer to any alternative possibility (which could lie in the future, or in the past). However, as the work presented here shows, it matters whether the imagined changes that lead to these alternative possibilities lie in the future or in the past. Pearl (2000) proposes a causal hierarchy in which knowledge on level I supports prediction, knowledge on level II supports hypothetical reasoning about the possible consequences of future interventions, and knowledge on level III supports counterfactual reasoning about how things could have turned out differently from how they actually did. Much work has demonstrated that the difference between levels I and II matters: correlation is different from causation (Bramley et al., 2017, 2018; Sloman & Lagnado, 2005; Steyvers et al., 2003). This study demonstrates that the difference between level II and III matters, too. Counterfactual simulation explains causal judgments, whereas hypothetical simulation doesn't. This result extends recent work showing that people differentiate between indicative conditionals (level I) and counterfactual conditionals (level III), and that causal judgments align more closely with counterfactual judgments (Skovgaard-Olsen et al., 2021).

The tight link between counterfactuals and causal judgments puts pressure on theories of causal judgment that don't distinguish between counterfactuals and other types of conditionals (e.g. Sebben & Ullrich, 2021). For example, the mental model theory analyzes causation in terms of temporally ordered sets of possibilities (Goldvarg & Johnson-Laird,

2001; Khemlani, Barbey, & Johnson-Laird, 2014). It defines "A causes B to occur" as "given A, B occurs", whereas "A enables B to occur" is defined as "given A, it is possible for B to occur". However, as we have seen above, for capturing causal judgments about particular events, indicative conditionals won't suffice. For instance, "Given A, B occurs" would be true if both A and B were the effects of some common cause C, $A \leftarrow C \rightarrow B$. But in this case, A doesn't cause B (even if A were to regularly precede B in time).

The results also put pressure on process theories of causation that aim to explain causal judgments without the use of counterfactuals (Dowe, 2000; Salmon, 1994). For example, Wolff's (2007) force dynamics theory of causation analyzes different causal expressions such as CAUSED, ENABLED, or PREVENTED in terms of configurations of force vectors that represent the forces that are at play at the time of interaction between cause and effect (see Beller, Bennett, & Gerstenberg, 2020, for a counterfactual account that captures people's use of different causal expressions). For A to have caused B, A's and B's force vectors need to have pointed in different directions, whereas for A to have enabled B, the force vectors need to have been aligned. However, in the video clips that I used, the way in which balls A and B interact with one another is identical in clips 1 through 4, and in clips 5 through 8. The actual process by which A brings about the outcome doesn't change. What does change is the initial and final position of the block which influences what would have happened in the relevant counterfactual situation. A process model of causation that doesn't incorporate counterfactuals has no way of producing a different verdict across these sets of clips, while people clearly do (see Wolff et al., 2010, for a process model that does incorporate some counterfactual machinery).

The results reported here also have implications for work on causal selection (Hesslow, 1988). In most situations, outcomes are the result of a multitude of contributing factors. However, people systematically choose to select some factors and not others as "the" cause of the outcome (Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Kahneman & Miller, 1986). What explains people's causal selections? Researchers have identified a number of factors that influence people's causal selections that include event normality, and the causal structure of the situation. For instance, when two events bring about an outcome conjunctively such that each event was necessary for the outcome to come about, people have a tendency to cite the abnormal event as the cause of the outcome rather than the normal event. In contrast, when two events combine disjunctively such that each individual event would have been sufficent for the outcome to come about, people tend to cite the normal event as the cause of the outcome rather than the abnormal event (Gerstenberg & Icard, 2020; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015).

A number of different accounts have been proposed that strive to explain people's causal selections (Hitchcock & Knobe, 2009; Livengood, Sytsma, & Rose, 2017; Quillien, 2020). One prominent view is that causal selections are influenced by counterfactuals (Gerstenberg & Stephan, 2021; Kominsky & Phillips, 2019). Accordingly, certain counterfactuals come to mind more easily than others, and this affects what events are selected as causes (Phillips, Luguri, & Knobe, 2015). Another related view suggests that people select those events as causes that would make for good points of intervention (Girotto, Legrenzi, & Rizzo, 1991; Hitchcock, 2012; Kirfel, Icard, & Gerstenberg, 2022; Lombrozo, 2010). While these theoretical account lead to similar predictions in many instances, they come apart in

scenarios like the one presented here. This is because what would have been good to do this time (counterfactual), need not necessarily be the thing one should do the next time (hypothetical). Future work needs to look at situations in which optimal interventions and counterfactuals come apart to better understand what drives causal selections.

The development of counterfactual reasoning

Counterfactual reasoning is an impressive cognitive feat. One has to take into account what actually happened, mentally travel back in time, consider an intervention on the course of events, imagine how things would have played out, and compare that to what actually happened. Maybe unsurprisingly, children appear to master this task relatively late by around five years of age (Beck & Guthrie, 2011; Beck & Riggs, 2014; Carey, Leahy, Redshaw, & Suddendorf, 2020; Kominsky et al., 2021; Koskuba, Gerstenberg, Gordon, Lagnado, & Schlottmann, 2018; McCormack, Ho, Gribben, O'Connor, & Hoerl, 2018; McCormack, O'Connor, Beck, & Feeney, 2016; Nyhout, Henke, & Ganea, 2019; Rafetseder, O'Brien, Leahy, & Perner, 2021). While earlier work claimed that three year old children can already reason correctly about counterfactuals (Harris, German, & Mills, 1996), later work argued that these early successes may have been false alarms. In one of the scenarios in Harris et al.'s (1996) study, Carol walks across a floor with dirty shoes. When asked "what if Carol had taken her shoes off – would the floor be dirty?" even three year old children answered correctly with 'no'. However, it's possible that children answered this question without running through a counterfactual simulation of what would have happened, and relying on basic conditional reasoning instead (Leahy, Rafetseder, & Perner, 2014). In general, if shoes are dirty the floor gets dirty, and if shoes are clean the floor stays clean. To tease apart counterfactual reasoning from basic conditional reasoning, Rafetseder, Schwitalla, and Perner (2013) added a second character who also walked across the floor with dirty shoes. Now the correct answer to the question of whether the floor would have been dirty if Carol had taken her shoes off is 'yes', because of the other child. Rafetseder et al. found that in this setting in which the outcome was causally overdetermined, even six year old children tended to get it wrong.

Recently, Nyhout and Ganea (2019) reported mature counterfactual reasoning in four and five year olds. In their experiments, children saw blocks being put on a box that had the potential to make the box light up. In one of the trials, a blue block and a green block are put on the box, and the box lights up. Children were asked "if she had not put the green one on the box, would the light still have been on?". While the question is clearly a counterfactual question, it seems possible for children to answer the question without considering a counterfactual. Instead, they might merely consider the hypothetical situation of just putting the blue block on the box and then try and simulate what would happen.

While the literature on the development of counterfactual reasoning has witnessed some false alarms, it's very likely that there have been some misses, too. Demonstrating counterfactual reasoning in young children is challenging because of the verbal processing demands. The question "Would ball B have gone into the gate if ball A hadn't been there?" is a mouthful. Other cognitive capacities, such as theory of mind (the ability to reason about other people's mental states), have been demonstrated in young children by replacing explicit verbal measures with implicit measures, such as where children are looking

on a screen (Low & Perner, 2012, but see also Kulke, von Duhn, Schneider, & Rakoczy, 2018). It's possible that children would be able to display the capacity for counterfactual reasoning much earlier than what has been shown so far, if verbal task demands could be circumvented (cf. Kominsky et al., 2021).

That said, the results reported here raise the bar for what's required to demonstrate successful counterfactual reasoning. To show that children are really simulating counterfactual rather than hypothetical situations, we need a setting like the one here in which counterfactuals and hypotheticals come apart. Using the metaphor of Pearl's (2000) ladder of causation: while early work may have misinterpreted level I reasoning as level III counterfactual reasoning, more work is required to make sure that we aren't misinterpreting level II reasoning as counterfactual reasoning either (Beck, Robinson, Carroll, & Apperly, 2006).

If, as I argue, counterfactual reasoning is indeed required to accurately answer causal questions in this setting, then experiments like the ones presented here would provide a new approach for demonstrating counterfactual reasoning in children. This approach doesn't rely on asking children any explicit counterfactual questions. Instead, to demonstrate counterfactual reasoning, it would appear to be sufficient to show that children's causal judgments aligned with those of adults.

What could have been better, and what would be good

The work presented here suggests that the process of counterfactual simulation is critical for understanding causal judgments. However, there are some theoretical and empirical limitations. On the theoretical side, there is a question about how exactly the distinction between hypotheticals and counterfactuals in Pearl's (2000) framework maps onto the distinctions drawn here in this paper. In Pearl's (2000) framework, the difference between hypotheticals and counterfactuals arises from the computational tasks that the model can solve at different levels of the hierarchy. Only on level III is the model able to first condition on what actually happened, and then consider an intervention that is counter to what actually happened. As illustrated in the sprinkler example (Figure 3), one way to implement this computationally is via using a twin network whereby the conditioning step and the intervention step are carried out one after the other in separate networks (see also Bareinboim et al., 2020). More generally, in order to be able to answer counterfactual questions, a model needs to know how the different variables functionally relate to one another (see the structural equations in Figure 3). In contrast, to answer hypothetical questions, it's sufficient to know the probabilistic contingencies between variables as well as their causal connections (knowing the structural equations isn't necessary).

In the experiments presented here, the difference between hypotheticals and counterfactuals isn't due to the model (or the participant) having causal knowledge at different levels of the hierarchy. Our participants know how physics works and they can simulate different possible scenarios.[7] Instead, the difference between hypotheticals and counterfac-

[7] In fact, the causal knowledge that people bring to bear on this task arguably goes beyond what's captured in Pearl's hierarchy. While structural equations are a very useful formal tool for representing causal knowledge, they don't naturally capture the kinds of spatio-temporal dynamics that are at play in these physical interactions (see Gerstenberg & Tenenbaum, 2017; Goodman, Tenenbaum, & Gerstenberg, 2015; Ullman, Spelke, Battaglia, & Tenenbaum, 2017).

tuals arises from how much information participants have about what happened. While the clip was paused shortly before the causal event of interest in the hypothetical condition of the experiment, in the other conditions, participants viewed the clip until the end. I highlighted that a key difference between hypotheticals and counterfactuals is whether the causal intervention takes place in the future or in the past. This difference in the time point of the intervention doesn't perfectly map onto the difference between level II and level III in Pearl's hierarchy. Nonetheless, the results reported here show that counterfactuals are critical for causal judgments. To explain the pattern of causal judgments across the different video clips, a model requires the causal knowledge and computational capacities on level III of Pearl's hierarchy.

On the empirical side, it will be important to document more broadly how counterfactual simulation and causal judgments relate. For example, I focused on a single setting here in which participants were asked to make causal judgments about physical events. Future work should expand this by assessing how counterfactuals and causal judgments are linked in a broader range of physical settings, as well as in settings that go beyond the physical domain (e.g. Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021). The experiments featured a relatively small set of clips. For a better quantitative assessment of how well the simulation model captures people's judgments, future work should include a larger set of test clips. In the setting here, there was a 50% that the block would move. Future work could manipulate this probability to see how it affects participants' judgments. The simulation model predicts that uncertainty about the block's movement should only affect hypothetical judgments but not the counterfactual, or causal judgments. That said, much work has shown that the (ab)normality of events influences causal judgments (Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Kominsky et al., 2015), and it's possible that such effects would be observed in this setting, too (see Gerstenberg & Icard, 2020; Kirfel et al., 2022). For example, consider a situation in which the block is initially out of the way but then moves in front of the gate. When ball A knocked ball B into the gate (via the wall and around the block), would the causal judgments be greater when there was a low chance that the block would move compared to when there was a high chance? Finally, it would also be interesting to use process tracing techniques, such as eye-tracking (Gerstenberg et al., 2017), to gain more direct evidence for where the mind travels when it makes causal judgments.

Acknowledgments

Thanks to Jingren Wang for help in the early stages of the project, to Ari Beller for help with implementing the simulation model, to David Rose for thoughtful comments on the manuscript, and to Thomas Icard for many fruitful discussions.

Appendix
Parameter search for the simulation model

The simulation model has two free parameters. One parameter determines how much the ball's velocity vector is rotated at each time step in the physical simulation. The degree of rotation is drawn from Gaussian distribution $\mathcal{N}(0, \sigma_{\text{ball}})$. The other parameter determines at what moment in time the block starts moving out of the way. The time point is determined by adding Gaussian noise with $\mathcal{N}(0, \sigma_{\text{block}})$ to the true moment in time at which the block moved. In clips 1-4, the block starts to move at time step 280 in the video clip, and in clips 5-8 it moves a time step 290. When determining at time step the block moves in the simulation, the model made sure that the block wouldn't begin moving before the two balls collided, and not after the end of the clip (the clip timed out at time step 700). When simulating what would happen in the hypothetical condition, the model first determines whether or not the block moves (it moves with 50% probability), and then determine at which point in time in starts moving.

Figure A1 shows the results of a grid search over the parameter space. For each parameter setting, I ran 1000 simulations for each of the eight clips in the hypothetical and the counterfactual condition. The best-fitting set of parameters is $\sigma_{\text{ball}} = 0.6$ and $\sigma_{\text{block}} = 175$. These parameters minimize the squared error between model predictions and participants' mean judgments for the different clips across both the hypothetical and counterfactual condition in Experiment 1. The model predictions shown in Figure 7 use these best-fitting parameters.

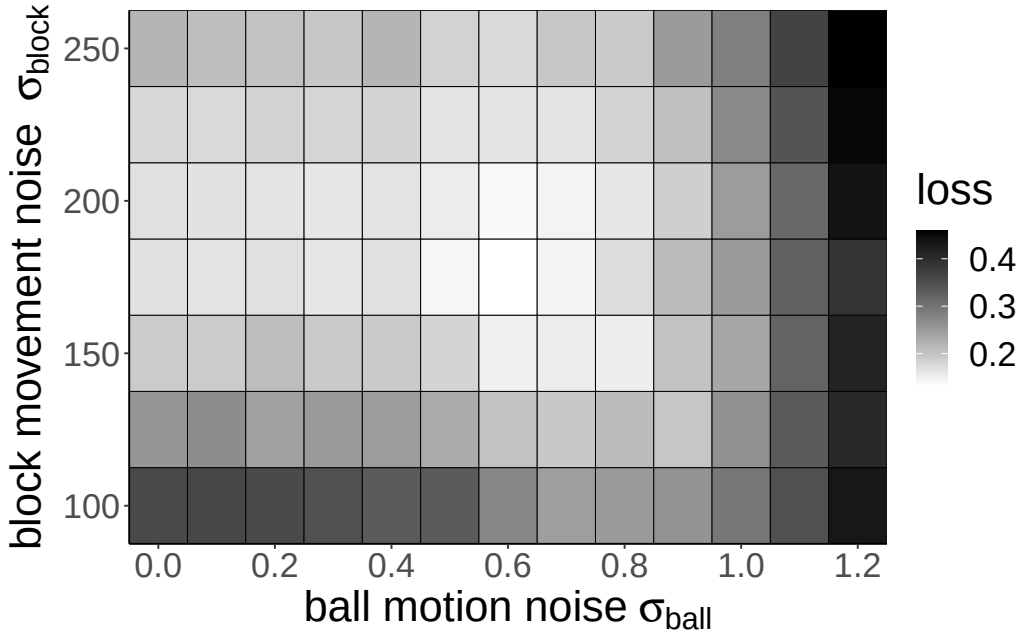

Figure A1. Results of a grid search over the ball motion noise parameter (σ_{ball}) and the block movement noise parameter (σ_{block}). The loss displayed here is the sum of squared errors between model predictions and participants' mean judgments (both on a scale from 0 to 1) for the 8 clips in the hypothetical and counterfactual condition (as shown in Figure 7).

References

Adams, E. (1965). The logic of conditionals. *Inquiry*, *8*(1-4), 166–197.

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.

Bareinboim, E., Correa, J., Ibeling, D., & Icard, T. (2020). On Pearl's hierarchy and the foundations of causal inference. *Sociological Methodology*, *40*(1), 75–149.

Beck, S. R., & Guthrie, C. (2011). Almost thinking counterfactually: Children's understanding of close counterfactuals. *Child development*, *82*(4), 1189–1198.

Beck, S. R., & Riggs, K. J. (2014). Developing thoughts about what might have been. *Child development perspectives*, *8*(3), 175–179.

Beck, S. R., Robinson, E. J., Carroll, D. J., & Apperly, I. A. (2006). Children's thinking about counterfactuals and future hypotheticals as possibilities. *Child Development*, *77*(2), 413–426.

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3133–3139). Cognitive Science Society.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, *124*(3), 301.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

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

Byrne, R. M. J. (2005). *The rational imagination: How people create alternatives to reality.* MIT Press.

Carey, S., Leahy, B., Redshaw, J., & Suddendorf, T. (2020). Could it be so? the cognitive science of possibility. *Trends in Cognitive Sciences*.

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

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

Ciardelli, I., Zhang, L., & Champollion, L. (2018). Two switches in the theory of counterfactuals. *Linguistics and Philosophy*, *41*(6), 577–621.

Dehghani, M., Iliev, R., & Kaufmann, S. (2012). Causal explanation and fact mutability in counterfactual reasoning. *Mind & Language*, *27*(1), 55–85.

Douven, I., & Verbrugge, S. (2010). The Adams family. *Cognition*, *117*(3), 302–318.

Dowe, P. (2000). *Physical causation.* Cambridge, England: Cambridge University Press.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2386–2391). Austin, TX: Cognitive Science Society.

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., & Icard, T. F. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

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

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*.

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.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1-3), 111–133.

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

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–653). MIT Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

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

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233–259.

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.

Hesslow, G. (1988). The problem of causal selection. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32). Brighton, UK: Harvester Press.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, *39*(4), 632–657.

Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, *93*(1), 75–88.

Hitchcock, C. (2001). A tale of two effects. *The Philosophical Review*, *110*(3), 361–396.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, *11*, 587–612.

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

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.

Kaufmann, S. (2013). Causal premise semantics. *Cognitive science*, *37*(6), 1136–1170.

Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience*, *8*.

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

Kominsky, J. F., Gerstenberg, T., Pelz, M., Sheskin, M., Singmann, H., Schulz, L., & Keil, F. C. (2021). The trajectory of counterfactual simulation in development. *Developmental Psychology*, *57*(2), 253–268.

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.

Koskuba, K., Gerstenberg, T., Gordon, H., Lagnado, D. A., & Schlottmann, A. (2018). What's fair? how children assign reward to members of teams with differing causal structures. *Cognition*, *177*, 234-248.

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic*, *10*(2), 201–216.

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? results from a systematic replication study. *Psychological science*, *29*(6), 888–900.

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.

Lassiter, D. (2017a). Complex antecedents and probabilities in causal counterfactuals. In *21st amsterdam colloquium* (pp. 45–54).

Lassiter, D. (2017b). Probabilistic language in indicative and counterfactual conditionals. In *Semantics and linguistic theory* (Vol. 27, pp. 525–546).

Leahy, B., Rafetseder, E., & Perner, J. (2014). Basic conditional reasoning: How children mimic counterfactual reasoning. *Studia logica*, *102*(4), 793–810.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, *85*, 297–315.

Livengood, J., Sytsma, J., & Rose, D. (2017). Following the fad: Folk attributions and theories of actual causation. *Review of Philosophy and Psychology*, *8*(2), 273–294.

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

Low, J., & Perner, J. (2012). Implicit and explicit theory of mind: State of the art. *British Journal of Developmental Psychology*, *30*(1), 1–13.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual

reasoning. *Psychological Review*, *122*(4), 700–734.

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

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of experimental child psychology*, *141*, 1–22.

McCormack, T., Ho, M., Gribben, C., O'Connor, E., & Hoerl, C. (2018). The development of counterfactual reasoning about doubly-determined events. *Cognitive Development*, *45*, 1–9.

McCormack, T., O'Connor, E., Beck, S., & Feeney, A. (2016). The development of regret and relief about the outcomes of risky decisions. *Journal of Experimental Child Psychology*, *148*, 1–19.

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: Rational and heuristic models of causal decision making. *Open Psychology Journal*, *3*, 119–135.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*(3), 249–264.

Nyhout, A., & Ganea, P. A. (2019). Mature counterfactual reasoning in 4- and 5-year-olds. *Cognition*, *183*, 57-66.

Nyhout, A., Henke, L., & Ganea, P. A. (2019). Children's counterfactual reasoning about causally overdetermined events. *Child Development*, *90*(2), 610–622.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. USA: Oxford University Press.

Over, D. E., & Evans, J. (2003). The probability of conditionals: The psychological evidence. *Mind & Language*, *18*(4), 340–358.

Over, D. E., Hadjichristidis, C., Evans, J. S. B., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive psychology*, *54*(1), 62–97.

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.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

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

Quillien, T. (2020). When do we think that x caused y? *Cognition*, *205*, 104410.

Rafetseder, E., O'Brien, C., Leahy, B., & Perner, J. (2021). Extended difficulties with counterfactuals persist in reasoning with false beliefs: Evidence for teleology-in-perspective. *Journal of Experimental Child Psychology*, *204*, 105058.

Rafetseder, E., Schwitalla, M., & Perner, J. (2013). Counterfactual reasoning: From childhood to adulthood. *Journal of Experimental Child Psychology*, *114*(3), 389–404.

Rips, L. J. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science*, *34*(2), 175–221.

Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–1135.

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

Schulz, K. (2011). "If you'd wiggled A, then B would've changed": Causality and counterfactual conditionals. *Synthese*, *179*(2), 239–251.

Sebben, S., & Ullrich, J. (2021). Can conditionals explain explanations? a modus ponens model of b because a. *Cognition*, *215*, 104812.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*(1), 1–51.

Skovgaard-Olsen, N., Stephan, S., & Waldmann, M. (2021). Conditionals and the hierarchy of causal queries. *Journal of Experimental Psychology: General*, *1*.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, *10*(9), 407–412.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247.

Sloman, S. A., & Lagnado, D. A. (2005). Do we 'do'? *Cognitive Science*, *29*(1), 5–39.

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*.

Stalnaker, R. C. (1970). Probability and conditionals. *Philosophy of science*, *37*(1), 64–80.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

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

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.

van Rooij, R., & Schulz, K. (2019). Conditionals, causality and conditional probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

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

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

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

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.

Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.

Yablo, S. (2002). De facto dependence. *The Journal of Philosophy*, *99*(3), 130–148.