

MIT Open Access Articles

*A counterfactual simulation model of
causal judgments for physical events.*

The MIT Faculty has made this article openly available. **Please share**
how this access benefits you. Your story matters.

Citation: Gerstenberg, Tobias, Goodman, Noah D, Lagnado, David A and Tenenbaum, Joshua B. 2021. "A counterfactual simulation model of causal judgments for physical events.." Psychological Review, 128 (5).

As Published: 10.1037/REV0000281

Publisher: American Psychological Association (APA)

Persistent URL: <https://hdl.handle.net/1721.1/138370>

Version: Original manuscript: author's manuscript prior to formal peer review

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



A counterfactual simulation model of causal judgments for physical events

Tobias Gerstenberg^{*}
Stanford University

Noah D. Goodman
Stanford University

David A. Lagnado
University College London

Joshua B. Tenenbaum
Massachusetts Institute of Technology

Abstract

How do people make causal judgments about physical events? We introduce the counterfactual simulation model (CSM) which predicts causal judgments in physical settings by comparing what actually happened with what would have happened in relevant counterfactual situations. The CSM postulates different aspects of causation that capture the extent to which a cause made a difference to *whether* and *how* the outcome occurred, and whether the cause was *sufficient* and *robust*. We test the CSM in several experiments in which participants make causal judgments about dynamic collision events. A preliminary study establishes a very close quantitative mapping between causal and counterfactual judgments. Experiment 1 demonstrates that counterfactuals are necessary for explaining causal judgments. Participants' judgments differed dramatically between pairs of situations in which what actually happened was identical, but where what would have happened differed. Experiment 2 features multiple candidate causes and shows that participants' judgments are sensitive to different aspects of causation. The CSM provides a better fit to participants' judgments than a heuristic model which uses features based on what actually happened. We discuss how the CSM can be used to model the semantics of different causal verbs, how it captures related concepts such as physical support, and how its predictions extend beyond the physical domain.

Keywords: causality; counterfactuals; mental simulation; intuitive physics.

To appear in *Psychological Review*

^{*}Corresponding author: Tobias Gerstenberg, Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, Email: gerstenberg@stanford.edu.

Introduction

The white billiard ball *caused* the black ball to go into the pocket. Joe suddenly turned around and walked back home *because* he realized that he forgot his wallet. The fall of Lehman brothers *is responsible for* the financial crisis. These sentences all make sense to us. They don't merely tell us *what* happened but also *why*. They explain events by pointing to their causes and reasons. The concept of causation is central to our understanding of the world, and to our understanding of each other (Pearl & Mackenzie, 2018; Sloman, 2005). It is the glue that holds the universe together (Hume, 1748/1975; Mackie, 1974).

So far, no unified account exists of how people make causal judgments. In philosophy, there is a vigorous debate about how to best analyze causation, and the philosophers' struggles of getting to grips with causation is reflected in a mixed bag of empirical findings in psychology about what factors people deem relevant when judging causation (Einhorn & Hogarth, 1986; Lagnado, Waldmann, Hagmayer, & Sloman, 2007). The difficulty of finding a unified theory of causation has led some to endorse a pluralistic view, postulating two or more fundamentally different concepts of causation (e.g. Cartwright, 1995, 2004; De Vreese, 2006; Godfrey-Smith, 2010; Hall, 2004).

In this paper, we develop the *counterfactual simulation model* (CSM) which provides a unified account of how people make causal judgments about physical events. The CSM draws from philosophical theories about the nature of causation (Beebe, Hitchcock, & Menzies, 2009; Paul & Hall, 2013), prior psychological work on causal judgment (Kahneman & Tversky, 1982; Wolff, 2007), as well as from recent developments in causal modeling (Halpern, 2016; Halpern & Pearl, 2005; Pearl, 2000). The model rests on the following three key assumptions: First, causal judgments about physical events are about difference-making (Woodward, 2003). Only things that made a difference are causes. Second, to understand causal judgments about specific events ("This stone broke the window.") rather than general causal relationships ("Stones break windows."), one needs to analyze difference-making in terms of counterfactual contrasts, comparing what actually happened with what would have happened in relevant counterfactual situations (Danks, 2017; Lipe, 1991). Third, causal judgments are sensitive to the different ways in which a candidate cause can make a difference to the outcome. For example, the cause can make a difference to *whether* the outcome occurred, or to *how* it came about (Glymour et al., 2010; Glynn, 2017; Hitchcock, 1996; Lewis, 2000; Schaffer, 2005). The CSM unifies existing theories of causation by showing how these different *aspects* of causation can be expressed in terms of counterfactual contrasts operating over the same causal model of the domain.

In principle, the CSM can be applied to a variety of domains. Our primary focus here is on dynamic collision events. This is a natural domain for studying causal judgments, where the ability to compute relevant counterfactuals rests on basic abilities for prediction and mental simulation in intuitive physics. In the General Discussion, we will lay out how the CSM generalizes beyond the physical domain.

The paper is organized as follows. First, we motivate the problem of causal judgment and discuss some of the major philosophical theories of causation. Then we summarize existing psychological work on causal judgment. Afterwards, we introduce the counterfactual simulation model. Several studies support the model's predictions, starting with a simple setting involving one candidate cause, and moving on to a more complex setting that fea-

tures multiple candidate causes. We conclude by discussing remaining challenges and future directions.

The philosophy of causation

Consider the following scenario: Billy throws a stone at a bottle. The stone hits the bottle, and the bottle shatters. Billy's throw (C) caused the bottle to shatter (E). But how can we justify this intuitive verdict? In philosophy, there are two major frameworks for analyzing causation: *process theories* and *dependence theories*. According to process theories, C was a cause of E if C and E were connected via a spatiotemporally continuous process that transferred some quantity such as physical force from C to E (Aronson, 1971; Dowe, 2000, 2001; Fair, 1979; Machamer, Darden, & Craver, 2000; Salmon, 1984, 1994; Waskan, 2011). In the example, Billy's throw caused the bottle to shatter because the physical force that Billy generated when accelerating the stone was transferred to the resting bottle and led to its destruction.

According to dependence theories of causation, C was a cause of E if E was dependent on C. The notion of dependence has been formalized in different ways. Some theories say that C was a cause of E if, if E was regularly followed by C in the past (Hume, 1748/1975), or if C raised the probability that E will happen (Suppes, 1970). Others capture dependence in terms of counterfactuals: C was a cause of E if E would not have happened in the absence of C (Lewis, 1973; Mackie, 1974). Interventionist theories specify these counterfactuals in terms of hypothetical interventions (Pearl, 2000; Woodward, 2003). Applied to our example, Billy's throw caused the bottle's shattering because the bottle would not have shattered if one had intervened in the actual course of events and made it such that Billy didn't throw the stone.

For simple scenarios like these, both process theories and dependence theories yield the same verdict. However, consider a modification of the scenario in which Billy (C_B) and Suzy (C_S) throw stones at the bottle (see Hall, 2004). Both their stones hit the bottle at exactly the same time and the bottle shatters. Each throw was such that it would have been individually sufficient to shatter the bottle. Was C_B a cause of the bottle's shattering? Intuitively, the answer is 'yes'. Both Billy and Suzy caused the bottle to shatter. Process theories have no trouble dealing with such situations of causal overdetermination: there was a spatiotemporally continuous process from each of Billy's and Suzy's throw, transferring force to the bottle (but see Ehring, 1986; Hitchcock, 1995).

Dependence theories, however, falter here (Halpern, 2016; Lewis, 1973). The bottle would still have shattered even if either Billy or Suzy hadn't thrown their rock. So, according to a simple counterfactual criterion, neither C_B nor C_S caused the bottle's shattering. One possible solution is to consider a different counterfactual contrast. According to Lewis's (2000) notion of *causal influence*, C is a cause of E if alterations to C would have resulted in alterations to E. This means that Billy's stone hitting the bottle is a cause of the bottle's shattering because if Billy's stone had hit the bottle slightly differently, then the bottle would have shattered differently (see also Woodward, 2011a). Whether C_B qualifies as a cause of E depends on what counterfactual contrasts for C_B and E are considered (see Gerstenberg & Stephan, 2020; Schaffer, 2005). In this example, the presence (versus absence) of C_B makes a difference to *how* the bottle shatters but not to *whether* it does.

The psychology of causal judgment

Much philosophical work on causation is concerned with determining what “a cause” is, and what distinguishes causal from a non-causal relationships. Psychologists, in contrast, are often more concerned with determining what people deem to be “the cause” of an outcome. What are the factors that make some causes seem more important than others? In this section, we review psychological work that has looked into how information about processes versus dependence affects people’s causal judgments. Most of this work has been qualitative but we will also discuss two recent formal modeling approaches, one rooted in process theories (Wolff, 2007), and the other in dependence theories of causation (Halpern & Pearl, 2005).

Empirical work on causal judgment. Information about covariation, counterfactual dependence, or causal processes affects people’s causal judgments (see Einhorn & Hogarth, 1986; Lagnado et al., 2007; Woodward, 2011b). Based on a comprehensive series of experiments with both adults and children from different cultures, Shultz (1982) concluded that people’s causal judgments are more consistent with process theories than with dependence theories of causation (Cheng, 1997; Cheng & Novick, 1992; Hume, 1748/1975). Shultz (1982) found that participants’ judgments were strongly affected by the presence of a plausible mechanism, and that dependence information, such as the timing of events, had little effect.

Mandel (2003) conducted a number of vignette-based studies in which he found that counterfactual and causal judgments sometimes come apart. When asked what would have needed to be different in order to undo an outcome, participants focused on events that were necessary for the outcome to occur, whereas for causal judgments, they focused on events that were sufficient under the circumstances (see Mandel & Lehman, 1998). While these results show that causal and counterfactual judgments can come apart, we believe that these judgments are nevertheless closely intertwined (see also Kominsky & Phillips, 2019). We will show that counterfactuals play a critical role in defining what it means for something to have been sufficient in the circumstances, as well as for analyzing *how* the outcome came about (see also Lombrozo, 2010; Walsh & Sloman, 2011).

So far we have seen evidence that people primarily care about process information when making causal judgments. However, sometimes counterfactual dependence appears to matter more. Chang (2009) directly pitted process theories and counterfactual theories against one another. In his experiments, a toy train runs into a card house causing the cards to fall. In some situations, an agent pushes the train such that the action is physically connected to the outcome. In other situations, the train is already moving and the agent opens a gate that would have otherwise blocked the train. While in both situations, the outcome was counterfactually dependent on the agent’s action, only the former situation involves a direct transmission of force between action and outcome. In order to manipulate counterfactual dependence, the outcome in some situations was overdetermined because of another train that approached the card house from the other side. For each situation, participants were asked to evaluate whether the agent’s action was a cause of the house of cards falling down. The results showed that participants’ causal judgments were most strongly influenced by counterfactual dependence. Participants gave significantly higher ratings when the house of cards wouldn’t have fallen but for the agent’s action. There was

no effect of physical connection on participants' judgments: whether the agent pushed the train or opened the gate didn't matter.

As we will see below, the CSM provides a unified account of these seemingly conflicting findings by postulating that people's causal judgments are sensitive to different aspects of causation. Some of these aspects relate more closely to causal processes, while others capture broader notions of dependence.

Theories of causal judgment. Research into causal judgments has suffered from a lack of formally specified models. The studies we have discussed so far have relied on comparing qualitatively whether causal judgments are influenced by information about processes and dependence. We will now discuss two formal models of causal judgment, one rooted in process theories, and the other in dependence theories.

Force dynamics model (FDM). According to Wolff's (2007) *force dynamics model* (FDM), causal events involve an interaction between two parties, an agent and a patient (cf. Talmy, 1988). The FDM defines causal expressions such as "caused", "prevented", "helped", and "despite" in terms of configurations of forces that characterize how agent (A) and patient (P) interact with respect to some endstate (E).

For example, the FDM predicts that an agent *caused* a patient to reach a certain endstate if the patient did not have a tendency to reach the endstate (i.e. P's force vector did not point toward E), the agent's force and the patient's force were not concordant (i.e. A's and P's force vectors did not point in the same direction), and the patient did in fact reach the endstate. To make this concrete, consider a situation in which a small boat (the patient) cruises on a pool with fans (the agent) located on the side of the pool. The boat is initially not headed toward a cone in the water (the endstate). However, at some point, the fans are turned on and the wind affects the boat in a way that it changes its direction and hits the cone. The force dynamics theory predicts that in this situation the fans *caused* the boat to hit the cone. If, in contrast, the boat was already headed toward the cone and the fans blew straight from behind, the FDM predicts that the fans *enabled* the boat to reach the cone. Wolff (2007) reports several experiments showing that the FDM accurately predicts participants' selection of causal terms across a variety of situations.

The FDM explains the use of different causal expressions as arising from a direct mapping between the force configurations and the causal terms, and without the need for any counterfactuals in the process. While the FDM supports counterfactual simulation (e.g. one can use a given force configuration to predict what would happen if the agent's or patient's force had been absent), counterfactuals don't feature in the definition of the different causal terms. Force configurations are primary, and both causal as well as counterfactual judgments derive from them.

Wolff, Barbey, and Hausknecht (2010) extended the FDM and incorporated counterfactuals to handle causation by omission as well as more complex causal interactions involving more than two participants. Causation by omission is generally difficult to accommodate by process theories of causation (Gerstenberg & Stephan, 2020; McGrath, 2005; Schaffer, 2000a). How can absences cause events when they clearly don't transfer any force? To deal with causation by omission, the FDM employs the concept of a "virtual force" – a force that would have been realized if something about the situation had been different. While Wolff et al. regard counterfactuals as being important for handling certain cases of causation by omission, they maintain that for assessing simple causal relations, counter-

factuals are not required. We believe, in contrast, that causal judgments are intimately linked to counterfactuals, and that even understanding simple causal judgments requires considering counterfactual contrasts.

Structural causal model (SCM). To discuss how causal judgments can be captured formally within the framework of dependence theories, we will focus on the *structural causal model* (SCM) developed by Halpern (2016) (for related accounts, see Hitchcock, 2001; Woodward, 2003; Yablo, 2002).

The SCM represents events of interest as variables, and the causal relationships between events are defined by structural equations relating the variables. Imagine that Billy (B) and Suzy (S) throw stones at a bottle. If either of them hits the bottle, the bottle shatters ($BS = B \vee S$). For simplicity, let's assume that each variable is binary so that, for example, Billy's throw can either hit the bottle ($B = 1$) or miss it ($B = 0$).

The relations between the variables capture how the world works. For example, the model expresses that if Billy's stone hits the bottle ($B = 1$), then the bottle shatters ($BS = 1$) no matter what Suzy does. However, the structural equations do not yet answer the question of whether one variable caused another in a particular situation. Was Suzy's hitting the bottle ($S = 1$) a cause of the bottle's shattering ($BS = 1$) in a situation in which Billy also hit the bottle ($BH = 1$)?

Much work has gone into defining the right criteria so that the model's causal verdicts agree with human intuition (Halpern & Pearl, 2005; Hitchcock, 2001; Woodward, 2003; Yablo, 2002). What all of these accounts have in common is that they take a simple test for counterfactual dependence as a starting point. Suzy's hitting the bottle was a cause of the bottle shattering if the bottle would not have shattered, had Suzy not hit it. While counterfactual dependence is sufficient for causation it is not necessary. The bottle would still have shattered even if Suzy hadn't hit it because Billy hit the bottle as well. Nevertheless, Suzy's hitting the bottle was clearly a cause of the bottle shattering.

To accommodate this intuition, the SCM defines causation so that one variable can qualify as a cause of another even if the two variables weren't counterfactually dependent in the actual situation, as long as there is a possible situation in which they would have been. For example, the bottle's shattering would have depended on Suzy's throw if Billy had missed the bottle. While this example is fairly straightforward, developing a definition of causation that agrees with people's intuitions across a range of different situations has proven challenging (see Halpern, 2016).

Bridging process and dependence accounts of causation. Both process and dependence accounts capture key aspects of how people make causal judgments (see Ney, 2009; Woodward, 2006). The counterfactual simulation model (CSM) aims to combine the best of both worlds (see also Strevens, 2013, for a unified perspective on causal explanation). In line with dependence theories of causation, we believe that people's causal judgments are fundamentally about difference-making. Only factors that made a difference in one way or another are causal candidates for having brought about the outcome. And since we focus on particular rather than general causal relationships, difference-making has to be expressed in terms of counterfactual contrasts (Collins, Hall, & Paul, 2004; Hiddleston, 2005; Hoerl, McCormack, & Beck, 2011; Jackson, 1977; Pearl, 2000; Woodward, 2003).

Traditionally, dependence theories have focused exclusively on a coarse kind of dependence that we will refer to as "whether-causation" (Ahn & Kalish, 2000; Mandel, 2003).

For example, the variables in structural models commonly represent the presence versus absence of events (Woodward, 2015). Process theories, in contrast, have focused on what we will call “how-causation” which captures a more fine-grained dependence between cause and effect. We will show that both of these aspects of causation are critical to understanding causal judgments.

The process of mental simulation plays a central role in our account (Craik, 1943; Hegarty, 2004). In line with Wolff et al. (2010, p. 215) we believe that “people simulate the processes that produce causal relationships rather than simply specifying the dependencies that hold between one event or state and another”. The CSM is a concrete implementation of this idea. It predicts that people use a mental model of the situation to simulate what would have happened in different counterfactual contingencies (Chater & Oaksford, 2013; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Kahneman & Tversky, 1982; Roese, 1997; Waskan, 2003). A detailed generative model of the situation allows one to express both whether a candidate cause made a difference to *whether* the outcome occurred as well as to *how* it came about (Jensen, 2019; Lewis, 2000; Woodward, 2011a).

In sum, the CSM unifies process and dependence theories by assuming that people represent their knowledge about how the world works as a generative model that captures the causal processes by which outcomes are produced (Gerstenberg & Goodman, 2012; Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008; Goodman, Tenenbaum, & Gerstenberg, 2015; Tenenbaum, Griffiths, & Niyogi, 2007). Instead of postulating fundamentally different *concepts* of causation that are associated with processes versus dependence (cf. Hall, 2004), the CSM posits different *aspects* of causation that are revealed through counterfactual contrasts on the generative model (see Williamson, 2006). These aspects express different ways in which a cause can make a difference to the outcome, such as *whether* or *how* it occurred. The CSM borrows heavily from Lewis’s (2000) idea of *causal influence*. However, whereas Lewis tried to reduce causation to fine-grained counterfactual dependence between cause and effect, the CSM maintains that causal judgments are sensitive to multiple aspects of causation that span across coarse-grained and fine-grained levels of dependence (see Woodward, 2011a).

Counterfactual Simulation Model (CSM)

The CSM makes predictions about the extent to which different candidate causes in physical settings are viewed as having caused a particular outcome to happen, or prevented it from happening. It does so in two steps: first, it uses a fine-grained test of difference-making to identify all candidate causes. It then uses additional counterfactual tests to predict the extent to which each identified candidate caused the outcome to happen. In other words, the first step filters out which candidates were “a cause” of the outcome. The second step, determines to what extent each of these was “the cause” of the outcome (Hart & Honoré, 1959/1985; Hesslow, 1988; Hilton, 1990).¹

Here, we illustrate how the CSM works by focusing on the task that participants faced in our experiments. In the General Discussion, we will discuss how the CSM may be used

¹Operating in these two steps increases the CSM’s efficiency in that some aspects of causation only need to be computed for a subset of the candidate causes. We don’t assume that people necessarily consider the different aspects of causation sequentially in that order.

to model causal judgments beyond this domain. In our experiments, participants viewed video clips of billiard ball collisions and judged whether one billiard ball caused another ball to go through a gate, or prevented that ball from going through. Figure 1a shows a diagrammatic illustration of one clip. In Experiment 2, participants saw more complex interactions between three billiard balls.

Much prior work has studied causal judgments by presenting participants with written vignettes (e.g. Alicke, 2000; Kahneman & Tversky, 1982; Lombrozo, 2010; Mandel, 2003; Walsh & Sloman, 2011). Vignettes are limited in that the counterfactuals need to be explicitly communicated. This way, one cannot be sure whether participants' causal judgments merely reflect demand effects, or whether they would have spontaneously sought out the relevant counterfactual information (see Gerstenberg, Peterson, et al., 2017). Our task provides a naturally graded, probabilistic sense of “whether-causation” due to the nature of the simulation mechanisms that people have available for predicting the outcomes of counterfactuals in this physical domain. It also provides a rich landscape of “how-causation” possibilities, because of all the different ways in which these collision events can unfold. Importantly, none of the existing models on causal judgments make quantitative predictions about people’s causal judgments in this task (including the forced dynamics model and the structural causal model discussed above).

Our use of video clips as experimental stimuli is inspired by a rich research tradition into the phenomenon of causal perception (Michotte, 1946/1963; Scholl & Tremoulet, 2000). Whereas work on causal perception focuses on the question of what factors make individual events look causal (e.g. whether the collision caused a ball’s movement, or whether the ball moved spontaneously), we focus on causal judgments about complex sequences of events that may involve multiple causes (e.g. judging to what extent different candidate causes are responsible for a target ball going through a gate). We will return to the question of how causal perception and causal judgments are related in the General Discussion.

Scope of the model

Before describing in detail how the CSM works, let us clarify the model’s scope. The CSM is a model of *causal judgment* and *not of causal learning* (Gerstenberg & Tenenbaum, 2017; Goodman, Ullman, & Tenenbaum, 2011; Kemp, Goodman, & Tenenbaum, 2010; Lake, Ullman, Tenenbaum, & Gershman, 2016; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Ullman, Goodman, & Tenenbaum, 2012; Wellman & Gelman, 1992). The CSM assumes that people already possess a causal model that incorporates the relevant domain knowledge for simulating different counterfactuals (see Bear et al., 2020; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018; Yi* et al., 2020, for work on how physical models may be learned). However, we don’t assume that people’s causal model is perfectly accurate, and we capture this uncertainty by introducing noise into the physical simulation model as described below (Smith et al., submitted; Ullman, Spelke, Battaglia, & Tenenbaum, 2017).

The CSM yields causal judgments about *particular events* (e.g. “Ball A caused ball B to go through the gate.”) rather than *general causal relationships* (e.g. “Ball A generally causes ball B to go through.”). Much prior work has focused on the question of how people infer the existence and strength of causal relationships between variables of interest (Cheng, 1997; Cheng & Novick, 1991; Griffiths & Tenenbaum, 2005; Jenkins & Ward, 1965). For example, participants might be asked to judge whether a drug was efficacious based on

information about the statistical contingency between patients who did and didn't received a drug and whether or not they were cured. The goal in these studies is to determine whether the statistical and temporal relationship between variables is merely coincidental or causal (Bramley, Gerstenberg, Mayrhofer, & Lagnado, 2018, 2019; Lagnado & Sloman, 2004, 2006). Instead, the CSM looks at unique singular events (e.g. "Ball B went through the gate."), and asks what caused these events to happen (Danks, 2017).

The CSM predicts causal judgments about *physical interactions*. We focus on the domain of dynamic collisions and assume that the observer has visual access to the relevant events. There are no unobserved causes in our setting, so the question of whether a particular event was caused by another, or merely happened by coincidence doesn't arise (see Griffiths & Tenenbaum, 2007, 2009). Even though the different aspects of causation that the CSM postulates are tailored to our task, we believe that these aspects are important beyond the physical domain and we will say more about that in the General Discussion.

The CSM makes predictions about the extent to which different *candidate objects* caused an outcome event. Most philosophical theories of causation take the causal relation to be events – cause events bring about effect events (see Menzies, 1989; Paul & Hall, 2013). For example, ball A's colliding with ball B is what caused ball B's going through the gate. Often, however, it is more natural to assign causal responsibility to an object (or an agent). We say that "a rock smashed the window" rather than "the collision between the rock and the window caused the window to smash" (see Croft, 1991; Pinker, 2007; Talmy, 1988; Thomason, 2014; Van Valin & Wilkins, 1996; Wolff, 2003). It is also often more natural to express counterfactual operations on candidate objects (e.g. what would have happened if ball A hadn't been present?) rather than on the events that they participated in (e.g. what would have happened if the balls hadn't collided?).

The CSM *is not a reductive account of causation* in the philosophical sense of reducing one concept to another one (cf. Lewis, 1973, 2000). For example, Lewis (1973) aimed to reduce causation to counterfactual dependence. He developed a possible-world semantics that determined the truth of counterfactuals by evaluating the similarity between different possible worlds with a similarity metric that was defined in non-causal terms. However, it has proven difficult to yield a satisfactory notion of similarity between possible worlds that does not itself rely on causal considerations. The CSM doesn't reduce causal judgments to a non-causal notion of counterfactual dependence. Instead the model assumes that general causal knowledge is required both for imagining counterfactual interventions (e.g. the removal of a candidate cause) and for simulating how the counterfactual situation would have unfolded (see Hiddleston, 2005; Pearl, 2000; Woodward, 2003). The CSM *is* reductive in the sense in that it explains causal judgments about particular events in terms of counterfactual operations defined over a general causal model of the domain.

The CSM yields *graded predictions*. There are two sources of gradation in the model's predictions. For one, people cannot know for certain what would have happened in relevant counterfactual situations. The CSM predicts that this uncertainty affects people's causal judgments. Another source of gradation comes from the fact that several aspects of causation jointly affect the overall causal judgment. The CSM explains interindividual differences by showing that participants' causal judgments are differentially affected by *whether* and *how* the candidate cause made a difference to the outcome.

Lastly, the CSM *doesn't solve the problem of causal selection*, that is, the problem of

deciding which causes are worth talking about in a given situation (Hesslow, 1988; Hilton, 1990). For example, people cite the striking of a match rather than the presence of oxygen as having caused a forest fire. Both event normality (Hitchcock & Knobe, 2009; Kahneman & Miller, 1986; Kahneman & Tversky, 1982) and the causal structure of the situation have been shown to influence people’s causal selections (Gerstenberg & Icard, 2019; Icard, Kominsky, & Knobe, 2017; Kirfel, Icard, & Gerstenberg, in prep; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015). In our experiments, we explicitly ask participants about the candidate causes, so the general problem of causal selection doesn’t arise. However, the CSM does address a narrower version of the problem in that it makes predictions about the extent to which different candidate causes are seen as causally responsible for bringing about the outcome.

Modeling counterfactual simulations in a probabilistic physics engine

The CSM assumes that people make causal judgments by simulating the outcomes of different counterfactual situations. Hence, it needs to give an account of counterfactual simulation. A key assumption of the CSM is that people have access to a generative model of the domain that supports running such mental simulations. There is growing evidence that people’s intuitive physical reasoning is based on approximate probabilistic inference in a mental physics engine that is in important ways analogous to the physics engines that are used in video games for generating realistic looking physical scenes (Ullman et al., 2017). The building blocks of a physics engine are objects with properties (such as shape, mass, friction, etc.) and approximately Newtonian mechanics that dictate how the objects interact with one another over time. This general approach for representing people’s intuitive physical understanding is extremely flexible and has been shown to explain people’s judgments in a variety of tasks that include making predictions about what will happen in the future (Battaglia, Hamrick, & Tenenbaum, 2013; Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016; Kubricht, Holyoak, & Lu, 2017; Smith et al., submitted; Smith & Vul, 2013), reasoning about what must have happened in the past (Gerstenberg, Siegel, & Tenenbaum, 2018; Smith & Vul, 2014), or inferring latent object properties such as mass or friction (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn, Mansinghka, & Griffiths, 2013; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015).

However, the physics engine itself cannot answer questions about causality. To do so, the CSM defines counterfactual operators. For example, to determine whether ball A caused ball B to go through the gate, the CSM compares what actually happened with simulations of what would have happened in the counterfactual situation in which ball A had been removed from the scene (or in which its initial position had been perturbed). So while the physics model encapsulates the general causal domain knowledge that dictates how objects move (or would have moved), the counterfactual operations on top of the physics engine are required to determine whether a candidate cause made a difference to the outcome of interest. For comparison, the *force dynamics model* (Wolff, 2007) uses force vectors to represent causal knowledge, and postulates that different causal expressions map onto different force configurations. The *structural equation model* (Halpern & Pearl, 2005) represents people’s causal knowledge with structural equations and yields causal verdicts about particular events via using a *do()*-operator (Pearl, 2000) for evaluating counterfactuals.

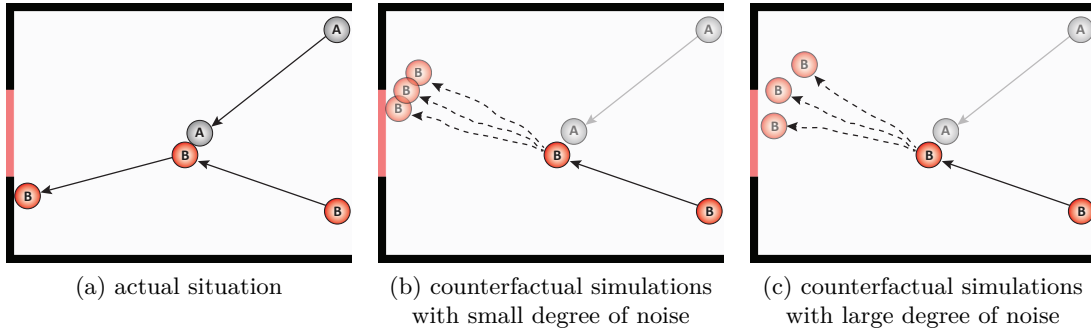


Figure 1. a) Illustration of what actually happened. b) and c) Illustrations of approximate simulations of where ball B would have ended up if ball A had been removed from the scene with small and large degrees of noise in the counterfactual simulations.

To assess what would have happened in a counterfactual situation in which the candidate cause had been absent, the CSM removes the cause from the scene and then simulates how things would have played out. Human observers don't have access to this ground truth – they only see what actually happened, not what would have happened. Different sources of uncertainty enter people's mental simulation of physical events. An observer may have perceptual uncertainty about the objects' positions in the scene, as well as dynamic uncertainty about how exactly the objects are going to move and interact with one another. For example, in the situation depicted in Figure 1a it is unclear whether ball B would have gone through the gate if ball A had been removed from the scene.

Physics engine-based models typically incorporate uncertainty by injecting small amounts of noise into the simulation dynamics for example by perturbing the position and velocities of objects in the simulator (Battaglia et al., 2013; Gerstenberg, Zhou, Smith, & Tenenbaum, 2017; Smith et al., submitted; Smith & Vul, 2013; Ullman et al., 2017). The same approach can be used to model uncertainty about counterfactual outcomes. The CSM models people's uncertainty in the counterfactual simulation by introducing noise to a ball's motion from the point at which the counterfactual situation diverges from the actual observed situation. In the case of Figure 1a, a random perturbation to the direction of ball B's velocity vector is introduced at each time step of the simulation after the time at which the collision with ball A would have occurred. A free parameter in the simulation model θ controls the standard deviation of the Gaussian distribution that determines the random perturbations that are applied to B's velocity vector at each time step in the simulation. Figure 1b shows three counterfactual simulations of where ball B would have ended up if ball A hadn't been present in the scene. Figure 1c shows three different counterfactual simulations where a larger degree of noise was added, representing a greater degree of uncertainty about what would have happened.

Causal connection: What was “a cause”?

The first question a model of causal judgment needs to answer is how to distinguish causes from non-causes. For process theories, a cause has to be connected to the effect via a spatiotemporally contiguous process. For dependence theories, the cause must have

made a difference to the outcome. Our proposed test for causal connection is inspired by both of these approaches. In line with dependence theories, we consider a counterfactual situation in which the candidate cause had been absent and evaluate whether the outcome would have been different in this case. In line with process theories, we assume that people use their understanding of the physical processes to mentally simulate what would have happened. We define the outcome event on a fine level of granularity that specifies not only whether or not the outcome happened, but also ‘when’ and ‘where’ it happened (see Paul, 2000; Woodward, 2011a).

Formally, we define an observer’s subjective degree of belief that a candidate cause C was a difference-maker (P_{DM}) of Δe as

$$P_{DM}(C \rightarrow \Delta e) = P(\Delta e' \neq \Delta e | S, \text{remove}(C)). \quad (1)$$

In words, to determine whether C was a difference-maker $P_{DM}(C \rightarrow \Delta e)$, the model first takes into account what happened in the actual situation S . A situation is defined by a full specification of the scene (e.g. the position of the walls and the gate) as well as the complete history of each ball’s motion path.

The model then considers a counterfactual situation in which the candidate cause had been removed from the scene $\text{remove}(C)$, and evaluates whether the outcome event in this counterfactual situation $\Delta e'$ would have been any different from the outcome event in the actual situation Δe . The Δ indicates that the outcome event is construed finely by including information about exactly where and when the outcome happened.² The $\text{remove}()$ operator is inspired by Pearl’s (2000) $do()$ operator and adapted to our domain of interest. Instead of implementing interventions by setting a variable to a particular value in a system of structural equations, the CSM intervenes in the physics engine that generated the observed clip by removing the candidate causal object. If a candidate cause C qualifies as a difference-maker then the model proceeds to evaluating the extent to which it caused the outcome.

In some situations, determining whether a candidate cause was a difference-maker is trivial. For example, whenever a candidate cause directly collided with the target, it qualifies as a difference-maker (see Figure 3a). In other situations, it is more difficult to assess whether the cause was a difference-maker. For example, the situation shown in Figure 3b shows a case of double prevention: ball B prevents ball A from preventing ball E from going through the gate. Here, ball B was a difference-maker of E’s going through the gate even though it didn’t collide with ball E. If ball B had been removed from the scene, than ball A would have knocked ball E out of the way.

In our experiments, we will look both at situations in which the target ball goes through the gate, and situations in which the target ball misses the gate, asking participants to judge to what extent the candidate prevented the target from going through the gate in

²We note that there are certain kinds of situations in which a candidate cause makes no difference to the outcome event (even when it is finely construed) but it intuitively nevertheless caused the outcome. For example, an earlier cause sometimes trumps a later cause in a way such that there would have been no difference to how the outcome had come about if either of the causes had been removed (Schaffer, 2000b). In order to deal with such cases, our model would need to be extended and allow for several causes to be removed at the same time when considering whether each of them qualifies as a cause of the outcome. For the domain we consider in our experiments, the problem of trumping causation does not arise.

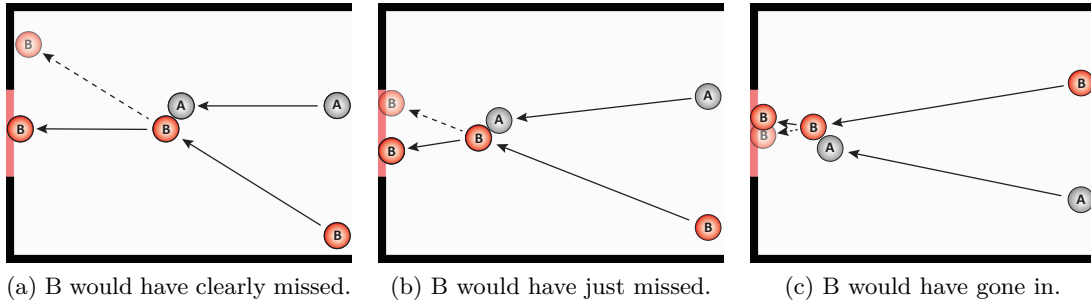


Figure 2. Schematic diagrams of collision events. Solid lines show the ball’s actual trajectories and the dashed line shows the trajectory ball B would have moved on if it hadn’t collided with ball A.

the latter case. For the CSM “prevented” is equivalent to “caused to miss” (but see Walsh & Sloman, 2011). Essentially, in the case of prevention e denotes the actual outcome of the target ball missing the gate, and e' denotes the counterfactual outcome of the ball going through the gate. In our descriptions of the model we will focus on situations in which the target ball goes through the gate but all of our studies feature an equal number of positive and negative outcomes.

Causal judgment: What was “the cause”?

Having identified a set of candidate causes as difference-makers, the CSM determines the extent to which each cause was “the” cause of the outcome. The CSM doesn’t assume that there is a fixed amount of causation to go around which would mean that multiple causes are necessarily in competition with one another. Instead, there could be multiple “good causes” or a single “bad cause” of an outcome. The CSM stipulates that people’s causal judgments are sensitive to four different aspects of causation, each of which is revealed through a different counterfactual test. We call the different aspects WHETHER-CAUSATION, HOW-CAUSATION, SUFFICIENT-CAUSATION, and ROBUST-CAUSATION, and discuss them now in turn.

whether-causation. Consider the three diagrams shown in Figure 2. In each diagram, the solid arrows indicate both balls’ motion paths before the collision, and ball B’s motion path after the collision. The dashed arrow indicates the motion path that ball B would have taken if ball A hadn’t been present in the scene. In all three situations, the two balls collided and B went through the gate. Since the two balls collided, ball A qualifies as a difference-maker P_{DM} of ball B’s going through the gate. But to what extent did ball A cause ball B to go through the gate in each case?

The CSM predicts that participants’ causal judgments are influenced by whether ball A’s presence made a difference to whether ball B went through the gate. We call this aspect whether-causation and define a person’s subjective degree of belief that a candidate cause C was a whether-cause P_W of outcome e as

$$P_W(C \rightarrow e) = P(e' \neq e | S, \text{remove}(C)). \quad (2)$$

Just like for difference-making, the model takes into account what happened in the actual

situation S , and then considers what would have happened in the counterfactual situation in which the candidate cause had been removed from the scene $remove(C)$. However, this time the outcome event is construed broadly. It only matters whether the outcome event happened or didn't happen (i.e. whether or not ball B went through the gate). C qualifies as a whether-cause of e to the extent that the observer believes that the outcome in the counterfactual situation in which C had been removed from the scene would have been (qualitatively) different ($e' \neq e$) from what it was in the actual situation.

Figure 2 shows three situations that differ in the extent to which ball A qualifies as a whether-cause. In Figure 2a $P_W(A \rightarrow e)$ is high. It is clear that ball B would not have gone through the gate if ball A had been removed from the scene. In Figure 2b, $P_W(A \rightarrow e)$ is intermediate. It is less clear what would have happened if ball A had been removed from the scene. Finally, in Figure 2c, $P_W(A \rightarrow e)$ is low. Ball B would have gone through the gate even if ball A had been removed from the scene.

In the experiments reported below, we use two complementary strategies for estimating the counterfactual probability of the target ball's going through the gate in the absence of the candidate cause. First, we ask human participants to judge whether they think the target ball would have gone through the gate if the candidate cause had been absent. Second, we model participants' counterfactual judgments as noisy simulations operating over their intuitive theory of the domain, as described above. To predict participants' counterfactual judgments, we draw samples from the approximate simulation model under different degrees of noise. For each sample, we record whether ball B would have gone through the gate, or would have missed the gate. We then use the proportion of samples in which ball B went through the gate to predict participants' judgment of whether ball B would have gone through the gate if ball A hadn't been present in the scene.

how-causation. Some counterfactual theories of causation try to capture people's causal judgments solely in terms of what we have termed whether-causation. Indeed, much of the empirical work discussed above has equated counterfactual theories of causation with a model that merely considers whether-causation, and then compared this simple counterfactual account with process models of causation that are more sensitive to the way in which the outcome actually came about. We believe that this dichotomy between process theories and counterfactual theories of causation is not helpful. From the research reported above, it is evident that people care about how events actually came about. However, this does not speak against counterfactual theories of causation. It merely suggests that only considering whether-causation is not sufficient for fully expressing people's causal intuitions. Counterfactual theories are flexible – they can express difference-making at different levels of granularity (Woodward, 2011a). The CSM uses a counterfactual test to determine whether a candidate cause made a difference to *how* the outcome came about.

Consider the diagram in Figure 3a. At the beginning of the clip, both the target ball E and one of the candidate causes, ball A, are stationary. Ball B, a second candidate cause, then enters the scene, hits ball A which consequently hits ball E, and E goes through the gate. To what extent do you think ball B caused ball E to go through the gate? What about ball A?

A counterfactual model that only considers whether-causation predicts the following in this case: Since both E and A are initially stationary, it is clear that E would not have gone through the gate if ball B had been removed from the scene. Thus, ball B is predicted

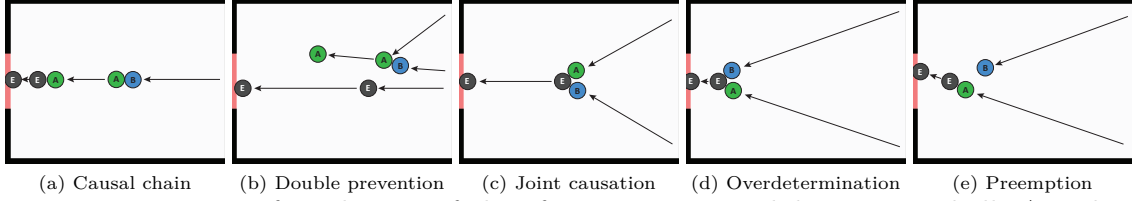


Figure 3. Diagrams of a selection of clips featuring two candidate causes, balls A and B, and one target, ball E.

to be seen as highly causal for E’s going through the gate. Ball A, in contrast, made no difference as to whether or not E went through the gate. Even if ball A had been removed from the scene, ball E would still have gone through the gate – it would have been knocked in by ball B. Thus, based on whether-causation only, we would predict that A has no causal responsibility for E’s going through the gate. However, there is clearly a sense in which ball A contributed to ball E’s going through the gate. Even though ball A’s presence did not make a difference as to *whether* ball E went through the gate, it clearly made a difference to *how* it did so.

How can one capture the intuition that ball A made a difference to ball E’s going through the gate? One part of the answer is that we need to construe the outcome event on a finer level of granularity just like we did for the test of difference-making. However, looking at the outcome event on a finer level of granularity is not enough. Instead, we need to introduce a different counterfactual test. When assessing how-causation, rather than considering what would have happened if the candidate cause had been *removed*, we consider what would have happened if the candidate cause had been *changed*. For example, to determine whether ball A was a how-cause of ball E’s going through the gate in Figure 3a the CSM simulates a counterfactual situation in which ball A’s initial position was somewhat perturbed and then records whether the outcome event would have been different on a fine level of granularity. If that’s the case, ball A qualifies as a how-cause of ball E’s going through the gate.

More formally, we define the probability that a candidate causal object C was a how-cause of event Δe as

$$P_H(C \rightarrow \Delta e) = P(\Delta e' \neq \Delta e | S, \text{change}(C)). \quad (3)$$

Taking into account what actually happened S , the CSM considers a situation in which the candidate cause was changed $\text{change}(C)$ and then simulates whether the event of interest in this situation would have been different from what it actually was $\Delta e' \neq \Delta e$. For the domain of dynamic collisions events considered here in this paper, we implemented the $\text{change}()$ operator as a very small perturbation to the initial position of a candidate cause. So if a ball actually was positioned at (x, y) at the beginning of the clip, we slightly changed that initial position to (x', y') . The outcome event Δe is construed finely and

includes information about exactly where and when the outcome happened.³

How-causation captures a key principle that motivates process theories of causation. It reveals whether there was a *transfer of force* from the candidate cause to the target (cf. Talmy, 1988; Wolff, 2007). This transfer of force can either be direct or indirect. For example, in the causal chain, ball A directly collides with ball E, whereas ball B only indirectly transfers force to ball E via ball A. Instead of requiring a conceptually distinct machinery (as introduced by process theories of causation), we provide a unified framework for expressing different aspects of causation in terms of counterfactual operations on a generative model (see Woodward, 2011a).

Note that while whether-causation can vary continuously between 0 and 1, how-causation is binary – a candidate cause either qualifies as a how-cause or it doesn't. In principle, more continuous notions of how-causation are possible. However, the simple binary notion suffices for our purposes. Note that even though the tests for difference making (Equation 1) and how-causation (Equation 3) are similar, they are not redundant. We will see below that a cause can be a difference-maker but fail to be a how-cause.

For whether-causation and how-causation, the CSM simulates the consequences of a counterfactual intervention on the candidate cause. By considering counterfactuals on alternative causes in the scene, the CSM captures two additional aspects of causation: sufficient-causation, and robust-causation.

sufficient-causation. Sufficiency is often discussed alongside necessity as a fundamental aspect of causation (e.g. Downing, Sternberg, & Ross, 1985; Hewstone & Jaspars, 1987; Icard et al., 2017; Jaspars, Hewstone, & Fincham, 1983; Mackie, 1974; Mandel, 2003; Pearl, 1999; Woodward, 2006). Ordinarily, necessity and sufficiency are defined on the level of general causal relationships (Cheng, 1997; Cheng & Novick, 1990, 1991; Jenkins & Ward, 1965). A cause is necessary if the effect never occurs in its absence, and sufficient if the effect always occurs in its presence. Because we are interested in people's causal judgments about particular events, we cannot use notions of necessity and sufficiency that are defined over repeated cause-effect contingencies.

Whether-causation captures necessity. A candidate cause was necessary if the effect would not have happened, had the cause been removed from the scene. Defining a notion of sufficiency for particular causal relationships is more involved. Previous proposals have in one way or another, relied on more general contingency information when defining sufficiency for particular events (see Cheng & Novick, 2005; Icard et al., 2017; Pearl, 1999; Stephan, Mayrhofer, & Waldmann, 2020; Woodward, 2006). Inspired by the structural-modeling account discussed above (Halpern, 2016; Halpern & Pearl, 2005) the CSM captures sufficiency by simulating a counterfactual situation in which all other candidate causes were removed from the scene, and then checking whether the candidate cause of interest would have made a difference to the outcome (broadly construed) in that situation.

³We acknowledge that the *change()* operator is somewhat vague. In fact, there are many different implementations that would yield the same result. For example, one could also consider a small perturbation to the ball's dynamic properties such as its mass, or initial velocity, or to a combination of these factors. One reason we chose this implementation is because it can be easily visualized (see Figure 6). More generally, how the *change()* operator is implemented depends on causal domain knowledge which dictates what properties have the potential of making a difference. For example, in our setup, a candidate ball's color doesn't make a difference to where and when an outcome of interest happened, but one could of course create situations in which color was causally relevant.

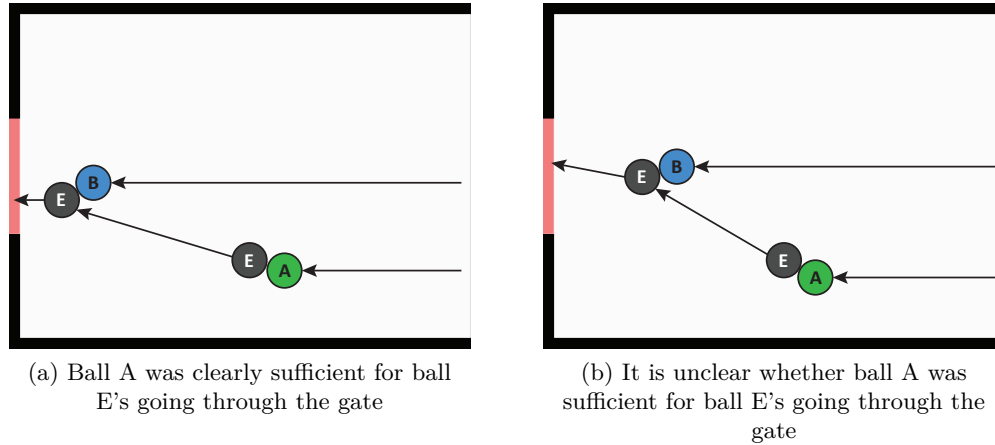


Figure 4. Two examples of clips in which an observer's subjective degree of belief differs that A was sufficient for E's going through the gate. In (a) it is clear that A was sufficient for E's going through the gate. In (b) it is less clear whether A was sufficient. E might have missed the gate in the counterfactual situation in which B had been removed from the scene.

More formally, the probability that a candidate cause C was a sufficient-cause of e is defined as

$$P_S(C \rightarrow e) = P_W(C \rightarrow e | \text{remove}(\setminus C)). \quad (4)$$

C is sufficient for e (broadly construed) if C would have been a whether-cause $P_W(C \rightarrow e)$ in a situation in which all other alternative causes had been removed $\text{remove}(\setminus C)$.⁴

To illustrate the sufficient-cause aspect, consider Figure 3c and Figure 3d. In both examples, ball E is initially at rest. Ball A and ball B hit ball E symmetrically (and simultaneously) such that ball E ends up going through the middle of the gate. In Figure 3c, both balls were necessary for ball E's going through the gate. If either of the balls had been removed from the scene, then ball E would not have gone through the gate. However, neither of the balls was individually sufficient. To check whether ball A was sufficient, the CSM considers a situation in which ball B had been removed, and checks whether ball A would have been a whether-cause in this situation. Since ball E would not have gone through the gate in if only ball A was present but not ball B, ball A was not sufficient for ball E's going through the gate. In contrast, in Figure 3d, neither of the balls were individually necessary for ball E's going through the gate. Ball E would have still gone through the gate even if either ball A or ball B had been removed from the scene. However, both ball A and ball B were individually sufficient for ball E's going through the gate. Ball A would have been a whether-cause in the counterfactual situation in which ball B had been removed from the scene (and vice versa).

⁴Note that the universal quantifier "all other alternative causes" is clearly too strong here. If this was applied to loosely, then nothing would qualify as a sufficient cause anymore – certain enabling conditions are almost always required to make an outcome happen (Mackie, 1974). So, a better way to think about the quantifier here is that it pertains to all *relevant* alternative causes. Our experiments sidestep the problem of how relevance is determined by explicitly stipulating what the alternative causes are.

In the same way in which an observer may be uncertain about whether a candidate cause was necessary for the outcome to occur, she may also be uncertain about sufficiency. Figure 4 shows two cases which differ in how clear it was that ball A was sufficient for ball E’s going through the gate. In both situations, ball E is initially at rest. Ball A and ball B collide with ball E and ball E goes into the gate. The clips differ in what would have happened if ball B had been removed. In Figure 4a it is relatively clear that ball E would have gone through the gate even if ball B hadn’t been present, and ball A was thus sufficient for ball E’s going through the gate. In contrast, in Figure 4b it is less clear whether ball E would have gone through the gate even if ball B hadn’t been present, so the probability that ball A was sufficient is lower here.

robust-causation. Both philosophers (Lewis, 1986b; Woodward, 2006) and psychologists (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Lombrozo, 2010; Vasilyeva, Blanchard, & Lombrozo, 2018) have argued that robustness is another important aspect of causal relationships. Causal relationships are robust to the extent that they would have continued to hold even if the background conditions had been different. We define robust-causation as

$$P_R(C \rightarrow e) = P_W(C \rightarrow e | \text{change}(\setminus C)). \quad (5)$$

Accordingly, C is a robust cause of e to the extent that the observer believes that C would have been a whether-cause in a counterfactual contingency in which all other candidate causes had been perturbed ($\text{change}(\setminus C)$). The *change* operator is the same here as the one introduced before to capture how-causation. We may consider different degrees of change to the alternative causes, perturbing their initial positions more or less. The more certain the model is that C would have been a whether-cause in a situation in which the alternative causes had been perturbed, the more robustly C brought about e . Robustness differentiates between cases in which a candidate cause directly brought about an outcome (like ball A in Figure 3e), from situations in which the causal relationship was mediated by other candidate causes (like in the causal chain in Figure 3a). For example, a golfer’s successful putt was robust if the ball would have gone in the hole even if other factors such as wind conditions had been varied. The putt was non-robust if it depended on a combination of factors, any of which, if they had changed slightly, would have led to the ball not going into the hole.

Putting it all together

The CSM predicts that people’s causal judgments are positively influenced by the extent to which the candidate cause was a whether-cause, a how-cause, a sufficient-cause, and a robust-cause of the effect event of interest. The CSM doesn’t commit to saying how much each aspect influences people’s judgments. People may differ in what aspects of causation they deem most relevant when judging causation. We use the term “causal responsibility” for the combination of aspects of causation and define the causal responsibility of a cause C for an outcome event e as

$$\text{Causal responsibility}(C \rightarrow e) = \alpha + P_{DM} \cdot (\beta_1 P_W + \beta_2 P_H + \beta_3 P_S + \beta_4 P_R). \quad (6)$$

Figure 5 illustrates the sequential nature in which the different counterfactual contrasts are considered. The model begins with a causal connection phase that determines for

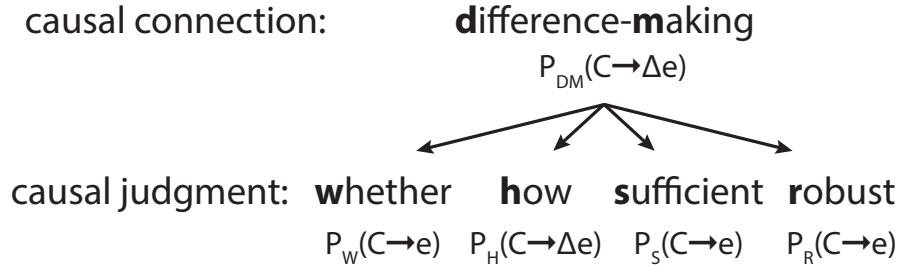


Figure 5. Relationship between the counterfactual contrasts. Difference-making is used as a criterion for whether a candidate qualifies as “a cause” of the outcome. The remaining aspects of causation (whether, how, sufficient, and robust) are used to determine the extent to which a candidate qualifies as “the cause” of the outcome. *Note:* Difference-making is computed first in the CSM because if $P_{DM}(C \rightarrow \Delta e) = 0$, none of the other aspects need to be computed (see Equation 6). However, the CSM doesn’t make the process-level prediction that these tests are computed in that sequence. For example, it’s possible that an observer may consider how-causation first.

each candidate cause whether it made a difference to the outcome. The difference-making test probabilistically selects amongst the candidates, the ones that were “a cause” of the outcome. For each identified candidate, the model then evaluates whether-causation, how-causation, sufficient-causation, and robust-causation to determine the extent to which each identified cause was “the cause” of the outcome. If a candidate cause doesn’t qualify as a difference-maker (i.e. $P_{DM}(C \rightarrow e) = 0$), then all of the other aspects of causation are zero because of the multiplication with P_{DM} . The β weights determine the extent to which each of the different aspects of causation influences a candidate’s causal responsibility for the outcome event, and the α is the intercept in the regression used for mapping between model predictions and participants’ response scale.

Figure 6 shows graphically, how the model evaluates the different aspects of causation. The different tests all have in common that they define a counterfactual operation over the physical representation of the scene, and compare the actual outcome with the outcome of a counterfactual situation. The tests differ in terms of what contingency they consider relevant, the counterfactual contrast, and the granularity at which the outcome event is specified.

1. **Relevant contingency:** The actual situation (whether & how), a situation in which all the other candidate causes were removed (sufficient), or in which they were changed (robust).
2. **Counterfactual contrast:** Either *removing* the candidate cause (whether, sufficient & robust) or *changing* it (how).
3. **Event granularity:** Either *coarse* whereby it only matters if the outcome happened or didn’t happen (whether, sufficient & robust), or *fine* (how) whereby it also matters where and when the outcome happened.

Let us illustrate how the full model works based on three of the example cases shown

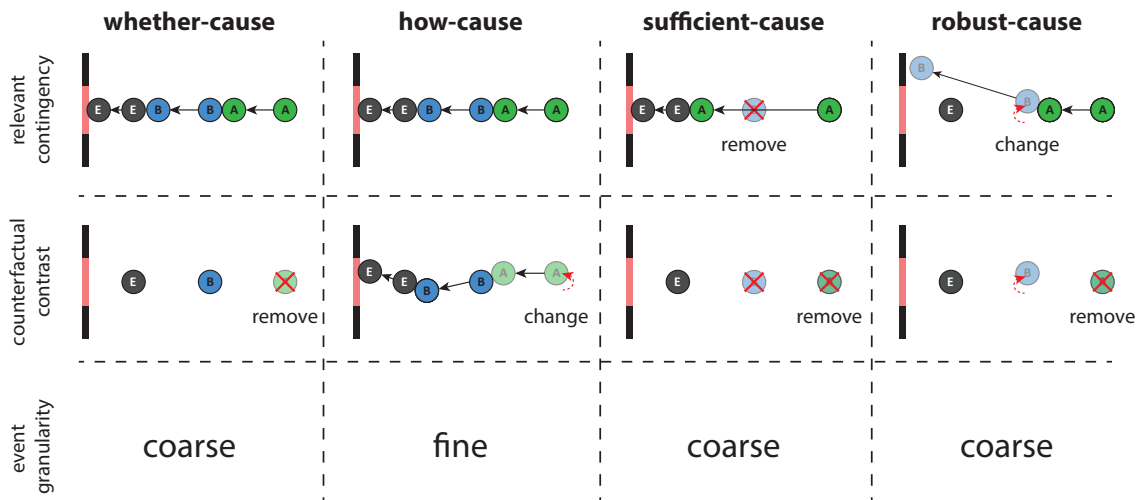


Figure 6. Illustration of the different types of counterfactual contrasts that are used to determine the extent to which ball A caused ball E to go through the gate in a causal chain. The top row shows the relevant contingency that serves as the starting point. This contingency is either the actual situation (for whether-causation and how-causation), or a situation in which the alternative cause (ball B) was removed from the scene (for sufficient causation), or its position randomly perturbed (for robust-causation). The middle row shows what counterfactual operation is considered. It either involves *removing* the candidate cause (ball A), or *changing* it. The bottom row shows at what level of granularity the outcome event in the actual situation and the counterfactual contrast are compared. At a coarse level of granularity, ball E either went through the gate or didn't. At a fine level of granularity, the “where” and “when” of E's going through the gate is considered.

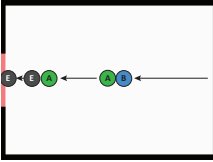
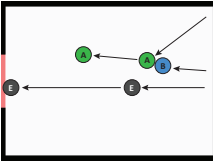
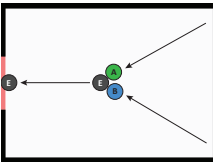
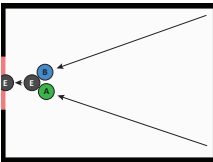
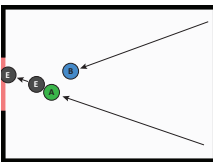
in Table 1. While the different aspects of causation vary continuously between 0 and 1 (with the exception of how-causation), for the sake of simplicity, we will treat each aspect of causation as either being true or false in these examples.

Causal chain. In the causal chain (Table 1, row 1), both ball A and ball B are difference-makers. Note that here, unlike in Figure 6, ball B is the first ball in the chain, and ball A is the second ball. If either ball had been removed from the scene, then the outcome event (finely construed) would have been different from what actually happened. Since both balls are difference-makers, the model proceeds to considering the other aspects of causation.

Ball A doesn't qualify as a whether-cause of ball E's going through the gate. Even if it had been removed from the scene, ball E would still have gone through the gate (because of ball B). However, ball A was a how-cause. If ball A's position was slightly perturbed, then the outcome event (finely construed) would have been different from what it actually was. Ball A was not a sufficient-cause for ball E's going through the gate. Ball A's presence would have made no difference to the outcome event in the counterfactual situation in which ball B had been removed from the scene. In a situation in which ball B is removed, ball E doesn't go through the gate no matter whether ball A is present or absent. Finally, ball A was not a robust-cause of ball E's going through the gate. There is only a small chance that

Table 1

Results of the different counterfactual tests applied to balls A and B. The predicted values for each aspect of causation are derived from running the counterfactual simulation model with the best-fitting noise parameter θ as described below.

Situation	Ball	$P_{DM}(C \rightarrow \Delta e)$	$P_W(C \rightarrow e)$	$P_H(C \rightarrow \Delta e)$	$P_S(C \rightarrow e)$	$P_R(C \rightarrow e)$
 causal chain	A	✓ (1)	✗ (0.34)	✓ (1)	✗ (0)	✗ (0.25)
	B	✓ (1)	✓ (1)	✓ (1)	✓ (0.67)	✓ (0.6)
 double prevention	A	✗ (0.05)	✗ (0)	✗ (0)	✗ (0)	✗ (0)
	B	✓ (0.91)	✓ (0.79)	✗ (0)	✗ (0)	✓ (0.72)
 joint causation	A	✓ (1)	✓ (0.88)	✓ (1)	✗ (0.12)	✓ (0.76)
	B	✓ (1)	✓ (0.89)	✓ (1)	✗ (0.11)	✓ (0.75)
 overdetermination	A	✓ (1)	✗ (0.01)	✓ (1)	✓ (0.99)	✗ (0.1)
	B	✓ (1)	✗ (0.01)	✓ (1)	✓ (1)	✗ (0.1)
 preemption	A	✓ (1)	✗ (0.23)	✓ (1)	✓ (1)	✗ (0.35)
	B	✗ (0)	✗ (0)	✗ (0)	✗ (0)	✗ (0)

Note: DM = difference-maker, W = whether-cause, H = how-cause, S = sufficient-cause, R = robust-cause. The values in parentheses show the quantitative predictions of the CSM. ✓ and ✗ indicate a value greater or less than 0.5, respectively.

ball A's being in the scene would have made a difference to ball E's going through the gate in a situation in which B's initial location was randomly perturbed. If ball B's position is randomly perturbed, then ball E will likely not go through the gate, no matter whether or not ball A is present.

Ball B was a whether-cause. Ball E would not have gone through the gate if ball B had been removed from the scene. Ball B was also a how-cause. The outcome event would have been slightly different, if ball B's initial position had been changed. Ball B was also a sufficient-cause. Ball B's presence would have made a difference to whether or not ball E ended up going through the gate in a counterfactual situation in which ball A had been removed from the scene. Ball B was a (somewhat) robust-cause of ball E's going through the gate. Even if A's position was slightly changed, ball B would still knock ball A into ball E.

Double prevention. In the double prevention case (Table 1, row 2), ball E goes through the middle of the gate on a direct path without making contact with either ball A or ball B. Ball A enters the scene in a way such that it would prevent ball E if nothing else happened. However, ball B knocks ball A out of the way. This is a case of double prevention since ball B prevents ball A from preventing ball E's going through the gate.

Ball A does not qualify as a difference-maker, so none of the other aspects of causation are considered. If ball A had been removed from the scene, then ball E would still have gone through the gate exactly in the same way as it did.

Ball B was as a difference-maker. If it had been removed then ball A would have knocked ball E out of the way. Ball B was a whether-cause but it wasn't a how cause. Even if ball B's initial position was changed, ball E would still have gone through the gate exactly in the same way that it did.⁵ Ball B was not a sufficient-cause. In a contingency in which ball A was removed from the scene, ball B would not have made a difference to the outcome. Ball B was a robust-cause of ball E's going through the gate. Even if ball A's initial position would have been somewhat perturbed, Ball B's presence would still have made a difference to whether or not ball E ended up going through the gate.

Preemption. In the preemption case (Table 1, row 5), ball E is initially at rest in front of the gate. Ball A collides with ball E, and ball E goes through the gate. Ball B enters the scene in a way such that it would knock ball E into the gate just a moment later. Thus, ball A preempts ball B from knocking ball E into the gate.

Ball A was a difference-maker. The outcome event (finely construed) would have occurred differently if ball A had been removed from the scene. Ball A was not a whether-cause. Ball E would have gone through the gate even if ball A had been removed from the scene (because of B). Ball A was a how-cause and a sufficient-cause. It would have made a difference to whether or not ball E ended up going through the gate in a counterfactual situation in which ball B had been removed from the scene. Ball A was not a robust-cause. Ball A's presence would not have made a difference to the outcome (broadly construed), if ball B's initial position had been somewhat perturbed.

Ball B was not a difference-maker. Ball E would have gone through the gate exactly in the same way in which it did even if ball B had been removed from the scene.

⁵There are of course ways in which one could change ball B's position such that ball E would *not* have gone through the gate. For example, one could change ball B's velocity so that it wouldn't collide with ball A anymore. However, when testing for how-causation, the model is constrained to considering small changes. There are many small changes to ball B that would not make any difference to the spatiotemporal details of ball E's going through the gate in the double-prevention case. For the causal chain, in contrast, any change to either of the candidate cause balls' positions would make a difference to the outcome event.

General information about experiments

We tested the CSM in one preliminary study, and two detailed experiments. Experiment 1 features one candidate cause, and Experiment 2 two. All studies were approved by the IRBs of the universities at which they were conducted (University College London and Massachusetts Institute of Technology). The complete materials including the video clips, diagrams, experiments, data, and analyses may be accessed here: <https://github.com/cicl-stanford/csm>

Description of the stimuli

The experiments were designed using Adobe Flash or Javascript. The videos were created using the flash and javascript implementation of the 2D physics engines Box2D (<http://box2d.org/>) and Chipmunk (<https://chipmunk-physics.net/>). Participants viewed collisions between balls from a bird's view perspective. Balls either entered the scene from the right or were present from the beginning and at rest. The scene was bounded by solid walls on the top, bottom, and left side, with a small gate in the middle on the left indicated by a red line. There was no friction and collisions were perfectly elastic without any loss of momentum. The experiment was presented in 800×600 pixels and the animations were updated at 30 frames per second. The scale from pixels to meter in the physics world was $\frac{1}{60}$ (i.e. the size of the stage was $13\frac{1}{3} \times 10$ m in the physics world). The radius of each ball was 0.5 m. In Experiment 1, some of the clips also featured a static rectangular brick $\frac{1}{4} \times \frac{5}{6}$ m or a teleport with a yellow rectangle as entrance ($\frac{1}{4} \times \frac{5}{6}$ m) and a blue circle as exit (radius = $\frac{1}{3}$ m). Clips in Experiment 1 featured a single collision event between balls A and B. The clips in Experiment 2 featured three balls and the number of collisions varied between clips.

Experimental design

Both experiments were run in two conditions: a counterfactual condition and a causal condition. In the *counterfactual condition*, participants were asked to judge whether they thought the target ball would have gone through the gate if the candidate ball hadn't been present in the scene. In the *causal condition*, participants were asked to evaluate the causal role of the ball(s) of interest.

Experimental procedure

The experiments were run online and participants were recruited via Amazon Mechanical Turk (Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2012). Only participants based in the US with an acceptance rate greater than 95% were allowed to participate in the experiments. Participants were paid at a rate of \$6 per hour. Each participant only participated in one experiment and no participants were excluded.

Preliminary study

In this preliminary study (see Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012), we wanted to test in a simple setting whether participants' cause and prevention judgments can be explained in terms of their subjective degree of belief that the presence of the

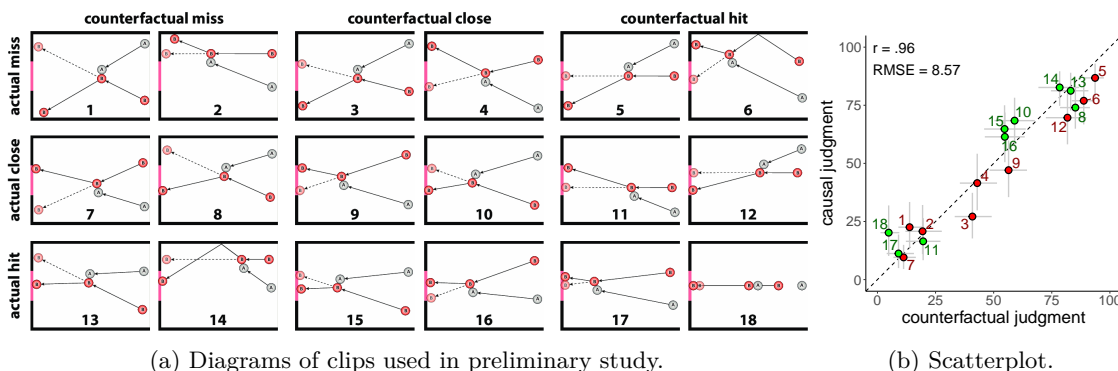


Figure 7. Preliminary study. (a) Diagrams of the 18 video clips that participants viewed in the study. The dashed arrows indicate how ball B would have moved if ball A hadn't been present. (b) Relationship between the average counterfactual judgments from one group of participants (x-axis) and the average causal judgments from another group of participants (y-axis). Each data point shows the averaged judgments for one clip. The labels next to the selection of clips correspond to the clip numbers in (a). For example, in clip 14 participants judged that ball A caused ball B to go through the gate, and they believed that ball B would have missed if ball A hadn't been there. In clip 2, participants judged that ball A didn't prevent ball B from going through the gate, and they believed that ball B would have missed the gate even if ball A hadn't been present. *Note:* The errorbars are 95% bootstrapped confidence intervals. green = situations in which ball B went through the gate, red = situations in which ball B didn't go through the gate.

candidate cause made a difference to whether or not the outcome occurred. The study featured 18 different clips shown in Figure 7a that varied whether ball B clearly missed the gate (“actual miss”), just missed the gate or barely went through (“actual close”), or clearly went through the gate (“actual hit”). The clips also varied what would have happened if ball A had not been present in the scene. Ball B would have either clearly missed the gate (“counterfactual miss”), just missed or barely gone through the gate (“counterfactual close”), or clearly gone through the gate (“counterfactual hit”).

The CSM predicts that participants' causal judgments will be dictated by what they believe would have happened in the relevant counterfactual situation. Because balls A and B collide in all of the clips, ball A trivially qualifies as a difference-maker and a how-cause of the outcome. Because ball A is the only candidate cause, the tests for sufficient-cause and robust-cause don't apply (these tests rely on considering contingencies in which alternative causes were removed but there are no alternative causes in this case). So, the CSM's predictions reduce to evaluating whether-causation for these clips (see Equation 2).

The study featured two different experimental conditions. In the *counterfactual judgment condition*, participants ($N = 41$) were asked to judge using a continuous slider whether the red ball (ball B) would have gone through the gate if the gray ball (ball A) hadn't been present. In the *causal judgment condition*, another group of participants ($N = 41$) was asked to judge whether ball A prevented ball B from going through the gate, whether it caused ball B to go through the gate, or did neither. Participants were instructed that they

could use intermediate values on the slider to express that ball A somewhat caused ball B to go through the gate, or somewhat prevented it from going through.

Figure 7b shows that there was a very close relationship between participants' counterfactual and causal judgments. The more participants believed that the outcome would have been different without ball A, the higher their causal judgment.⁶ For example, in clip 18 both balls initially travel toward the gate and ball A collides with ball B shortly before ball B enters the gate. Here, participants were certain that ball B would have gone through the gate even if ball A hadn't been present, and thus gave a low causal rating. In clip 16, it was less clear what would have happened in the relevant counterfactual situation, so participants gave an intermediate causal rating. In clip 14, it was clear that ball B would have missed, so participants judged that ball A caused ball B to go through the gate. Similarly, for situations in which ball B actually missed the gate, participants' prevention judgments increased the more certain they were that ball B would have gone through the gate without ball A (see clips 2, 4, and 6).

While the tight quantitative fit between counterfactual and causal judgments in the preliminary study supports the role of counterfactual simulation in people's causal judgments, the results do not rule out alternative explanations. As apparent from Figure 7a the set of clips varied what actually happened as well as what would have happened in the relevant counterfactual situations. Since all clips differed in what actually happened, it may be possible in principle to develop an account of the results by appealing to differences solely in what actually happened and without the need to rely on counterfactual contrasts. To address this concern, Experiment 1 provides a much stronger test of the CSM's prediction that causal judgments about physical events are inextricably linked to counterfactual simulation.

Experiment 1: One candidate cause

Like in the preliminary study, this experiment tests the CSM in a setting with a single candidate cause. However, now the stimuli were designed to ensure that no possible process theory can make the same predictions as the CSM. Because process theories focus only on what actually happened, in contrast to any kind of counterfactual analysis, we constructed pairs of clips in which the exact same events actually happen but different counterfactual trajectories obtain, and tested whether people judged causation differently in these pairs. Consider the pair of clips shown in Figure 8a and 8b. In both clips, the paths that balls A and B take are identical. What differs between the clips is what would have happened in the counterfactual situation in which ball A had been removed from the scene. In Figure 8a, ball B would not have gone through the gate even if ball A had been removed because it would have been blocked by the brick. In Figure 8b the brick's location is different. Here, ball B *would* have gone through the gate in the relevant counterfactual situation. If participants' causal judgments differ between these situations then this cannot be explained by what actually happened which was identical in both clips.

⁶Since we asked participants in the counterfactual judgment condition whether ball B would have gone through the gate if ball A had not been present in the scene, we can directly take their counterfactual judgments to predict participants' prevention judgments. To predict participants' causal judgments, we subtracted the counterfactual judgments from 100 to get the probability that ball B would have missed the gate in the absence of ball A.

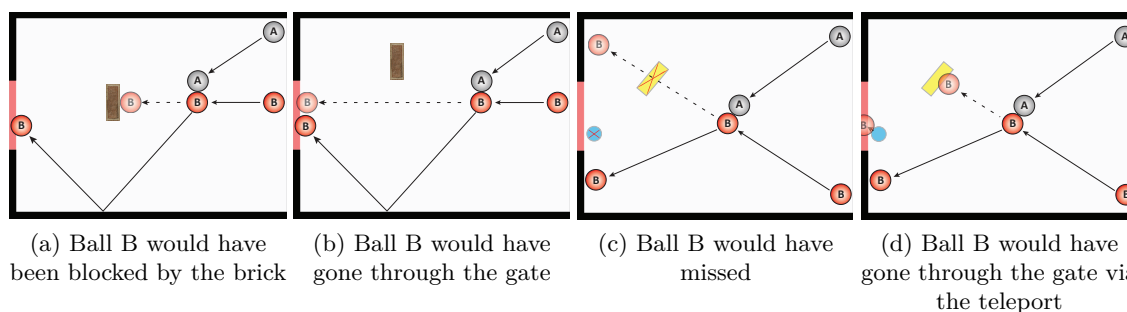


Figure 8. Illustration of two pairs of clips in which the interaction between balls A and B is identical, but the relevant counterfactuals differ. For each pair, there is one clip for which the counterfactual outcome (broadly construed) would have been different if ball A had been removed from the scene (a and c), and one clip for which the outcome would have been the same (b and d).

We also generated a number of clips in which we manipulated the counterfactual outcome without changing the spatial location of any components in the scene. For this purpose, we introduced participants to a teleport. The teleport only affects ball B but not ball A. It has an entry (the yellow rectangle) and an exit (the blue circle). Ball B exits the teleport in the same direction in which it enters it. Consider the pair of clips shown in Figure 8c and 8d. Again, what actually happened was identical in both clips. Because the teleport was switched off in Figure 8c (as indicated by the red cross on top of the teleport entry and exit), ball B would have missed the gate even if ball A had been removed from the scene. In contrast, when the teleport was on (Figure 8d), ball B would have gone through the gate via the teleport, if ball A had not been present in the scene. The teleport allows us to create situations in which the relevant counterfactual situation is different but what actually happened is identical including the spatial location of all the objects on the screen. The teleport further provides a test for the flexibility of people’s mental simulations.

Figure 9 shows diagrams of the 18 test clips that participants viewed. Like in the preliminary study, we varied the closeness of what actually happened (actual miss, actual close, actual hit), as well as how close the outcome would have been in the relevant counterfactual situation (counterfactual miss, counterfactual close, counterfactual hit). For each combination of actual and counterfactual closeness, we created two different clips (e.g. in clip 1 and 2, ball B clearly missed the gate in the actual situation, and it would have clearly missed in the counterfactual situation without ball A). For the cases in which the outcome was close (or would have been close), ball B went through the gate for half of the clips and missed the gate in the other half.

Predictions

Like in the preliminary study, the CSM predicts a high correlation between participants’ counterfactual and causal judgments. The more certain participants are that the outcome would have been different from what actually happened had ball A been removed (see the columns in Figure 9), the more they will say that ball A caused ball B to go through the gate (if it went through), or prevented ball B from going through the gate (if it missed).

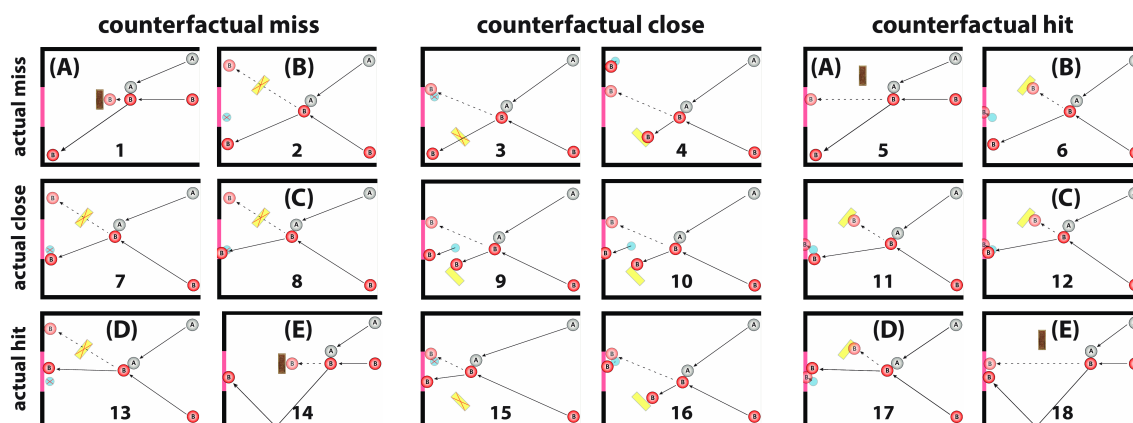


Figure 9. Diagrams of clips used in Experiment 1. *Note:* The solid arrows indicate the actual paths on which balls A and B moved before the collision and the path on which ball B moved after the collision. The dashed line indicates the path on which ball B would have moved if ball A had not been present in the scene. The numbers indicate the clip number. The letters indicate pairs of clips that are matched in terms of what actually happened. The brown rectangle represents a solid brick. The yellow rectangle and the blue circle represent the entry and exit of a teleport. A red cross on the teleport indicates that it’s switched off.

How close the actual outcome was should not affect participants’ causal judgments according to the CSM. It just matters whether or not ball B actually went through the gate. Importantly, the CSM predicts that participants’ judgments will differ substantially between clips in which what actually happened was identical but for which the counterfactual outcome would have been different (as indicated by the letters (A) - (E) in Figure 9). For example, the CSM predicts a high causal rating in 14 (E) where ball B would have been blocked by the brick, and a low causal rating in 18 (E) where ball B would have gone through the gate even if ball A hadn’t been present in the scene.

Methods

Participants and Procedure. 82 participants ($M = 35$ years, $SD = 12.2$, 45 female) participated in the experiment. Participants were instructed that they would see 20 different video clips in total (which included two practice clips shown first). In this experiment, we had each participant provide both counterfactual judgments as well as causal judgments. The order of the two judgment blocks was counterbalanced. The order of the 18 test clips was randomized in each of the two judgment blocks. On average, the experiment took $M = 21.25$ minutes ($SD = 5.11$) to complete.

In the *causal judgment block*, participants saw each clip played twice until the end. They were then asked on a separate screen to answer the question: “What role did ball A play?” Participants indicated their response on a sliding scale whose endpoints were labeled “it prevented B from going through the hole” (−100) and “it caused B to go through the hole” (100). The midpoint was labeled “neither” (0). Participants were instructed that they could use intermediate values on the slider to express that ball A somewhat caused ball B to go through the hole or somewhat prevented it from going through the hole. The

instructions in this condition did not mention anything about counterfactuals.

In the *counterfactual judgment block* the clips were paused shortly after the time at which the balls collided. Upon having viewed the clip for a second time, this question appeared at the bottom of the clip: “Would the red ball have gone through the gate if the gray ball had not been present?”. Participants provided their answers on a sliding scale from 0 (“definitely no”) to 100 (“definitely yes”). The slider was initiated at the midpoint which was labeled “uncertain”. After having indicated their response, participants received feedback by viewing the same clip again from the beginning with ball A removed from the scene.

At the beginning of the experiment, participants did not know that they will be asked to make both counterfactual and causal judgments. For example, participants who first made counterfactual judgments did not know that they will be asked to make causal judgments later on and vice versa.

Design. The experiment followed a 3 (*actual outcome closeness*: clear hit, close hit, clear miss) \times 3 (*counterfactual outcome closeness*: clear hit, close hit, clear miss) \times 2 (*question order*: causal before counterfactual, counterfactual before causal) design, whereby participants saw two different clips for each combination of actual and counterfactual outcome closeness (see Figure 9).

Results and Discussion

We will discuss participants’ counterfactual judgments first and then look at their causal judgments.

Counterfactual judgments. We first tested whether participants’ counterfactual judgments were affected by the question order (i.e. whether participants made counterfactual or causal judgments in the first block of the experiment). To do so, we fit two Bayesian mixed effects regression models.⁷ The full model contained actual outcome closeness, counterfactual outcome closeness, and question order as well as interactions between question order and actual/counterfactual outcome closeness as fixed effects, and random intercepts and slopes for actual and counterfactual outcome closeness. The reduced model only contained actual and counterfactual outcome closeness as fixed effects with the same random effects structure as the full model. We compared the two models using approximate leave-one-out cross-validation (PSIS-LOO; see Vehtari, Gelman, & Gabry, 2017) and found that the reduced model performs better (difference in expected log predictive density (elpd) of -3.5 with a standard error of 0.7) suggesting that question order didn’t affect participants’ counterfactual judgments. Because there was no effect of question order, we combined the counterfactual judgments from both question orders. Figure 10 shows participants’ mean counterfactual judgments for the different clips together with the predictions of the best-fitting approximate simulation model.

Recall that the approximate simulation model introduces some noise to capture people’s uncertainty about what would have happened if the candidate cause had been removed. This noise is introduced in the form of random perturbations to the direction of ball B’s motion at each time step in the simulation after the point at which the two balls would have

⁷All Bayesian models were written in Stan (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner, 2017) in R (R Core Team, 2019).

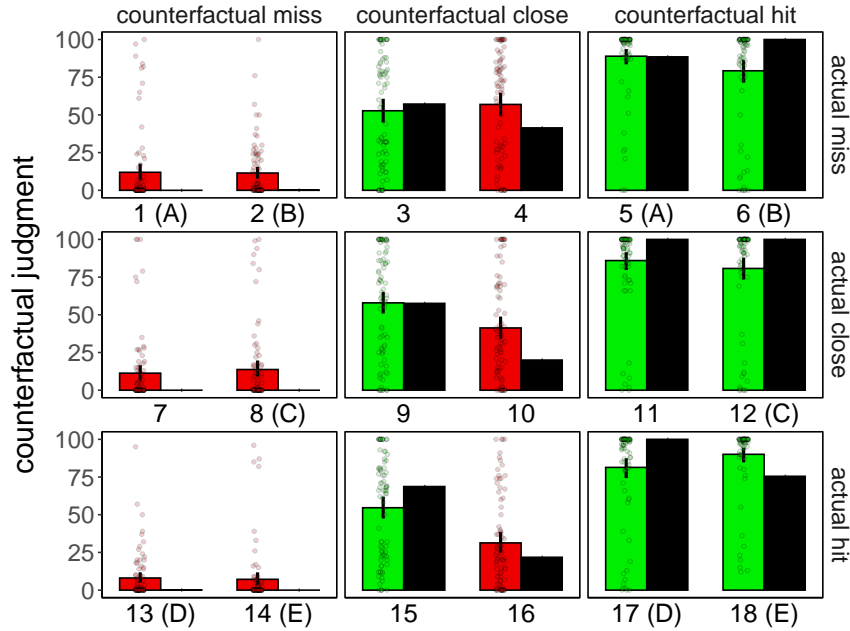


Figure 10. Mean counterfactual judgments (colored bars) together with the model predictions (black bars) of the best-fitting approximate simulation model. *Note:* Red bars indicate cases in which ball B would have missed. Green bars indicate cases in which B would have gone through the gate. Small points are individual judgments (jittered along the x-axis to increase visibility). Error bars are bootstrapped 95% confidence intervals.

collided. We generated model predictions by drawing 1000 noisy samples for each of the 18 clips using different degrees of noise ranging from $\theta = 0^\circ$ to $\theta = 2^\circ$ in steps of 0.1° , where θ refers to the standard deviation of the Gaussian distribution from which the perturbations to B’s velocity vector were drawn.

The black bars in Figure 10 show the proportion of cases in which ball B went through the gate out of the sample of cases that was generated for each clip. For example, in clip 3, ball B ended up going through the gate in 572 out of 1000 cases. We fit the noise parameter to participants’ judgments by minimizing the sum of squared errors between model prediction and participants’ mean counterfactual judgments. The best-fitting approximate simulation model has a noise value of $\theta = 0.9^\circ$ with $r = .96$ and $\text{RMSE} = 13.46$. For comparison, a deterministic model (i.e. $\theta = 0^\circ$) performs worse with $r = .86$ and $\text{RMSE} = 28.15$. This model simply predicts a rating of 0 for cases in which ball B would have missed the gate (the red bars in Figure 10) and a rating of 100 for the cases in which ball B would have gone in (the green bars).

Causal judgments. To check whether participants’ causal judgments were affected by question order, we conducted the same analysis as before with the counterfactual judgments comparing a full model that includes question order as a predictor with a reduced model that omits question order (and its interactions). The results of the approximate leave-one-out cross-validation revealed that the full model performs better (difference in expected log predictive density (elpd) of -3.2 with a standard error of 3.8), suggesting that participants’ causal judgments were affected by question order.

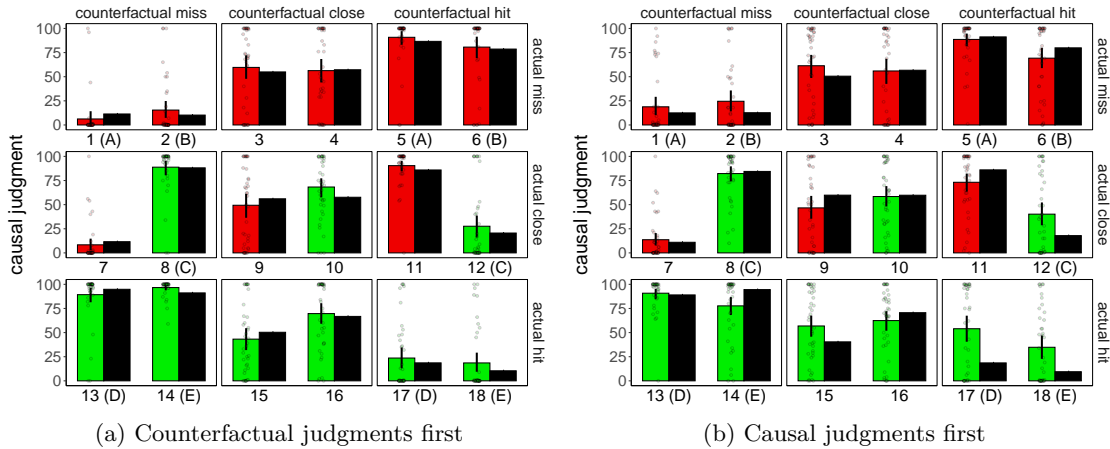


Figure 11. Causal judgments (colored bars) with model predictions (black bars). The model predictions are based on the mean counterfactual judgments in each condition. *Note:* Red bars are prevention judgments and green bars are causation judgments. Small points are individual judgments (jittered along the x-axis to increase visibility). Error bars are bootstrapped 95% confidence intervals.

Figure 11 shows the causal judgments separated by whether participants first made counterfactual judgments (Figure 11a) or causal judgments (Figure 11b). We used participants’ counterfactual judgments to determine P_W for each clip. Since we asked participants in the counterfactual judgment condition whether ball B would have gone through the gate if ball A had not been present in the scene, we can directly take their counterfactual judgments to predict participants’ prevention judgments. To predict participants’ causal judgments, we subtract the counterfactual judgments from 100 to get the probability that ball B would have missed the gate in the absence of ball A.

For those participants who answered the counterfactual questions first, the correlation between their counterfactual and causal judgments was very high with $r = .99$ and $RMSE = 5.27$. For the other group of participants who made causal judgments in the first block and counterfactual judgments in the second block, the correlation was somewhat lower but still high with $r = .92$ and $RMSE = 13.78$. Combining the causal judgments from both groups, a model that is based on participants’ mean counterfactual judgments correlates well with $r = .95$ and $RMSE = 10.43$, as does a model that is based on the best-fitting approximate simulation model (whereby $\theta = 0.9^\circ$ which was fit to participants’ counterfactual judgments) with $r = .92$ and $RMSE = 19.09$.

As predicted, participants’ causal judgments varied as a function of their subjective degree of belief that ball A’s presence made a difference as to whether or not ball B went through the gate (see Equation 2). Both cause and prevention judgments increased the more certain participants were that ball A’s presence was necessary for the outcome. This effect can be seen by comparing cause and prevention judgments between different columns in Figures 11a and 11b. How close ball B actually ended up going through the gate had little to no effect on participants’ judgments (compare different rows in which the outcome is the same in Figures 11a and 11b). For example, participants’ causal judgments didn’t

differ between clips 13 and 14 in which ball B went through the middle of the gate versus clip 8 in which ball B only barely went in. Similarly, participants' prevention judgments didn't differ between clips 5 and 6 in which ball B clearly missed the gate versus clip 11 in which ball B almost when through the gate.

The results of this experiment clearly demonstrate the importance of counterfactual contrasts for people's causal judgments. Participants' causal judgments differed strongly between clips in which what actually happened was held constant. For example, in clip 1, participants did not think that ball A prevented ball B from going through the gate. In this clip, the brick would have blocked ball B even if ball A had been removed from the scene. In contrast, in clip 5, participants judged that ball A prevented ball B from going through the gate. Here, the brick wouldn't have blocked ball B. Thus, even though the way in which balls A and B moved and collided was exactly the same in clips 1 and 5, participants' prevention judgments differed as a function of where the brick was placed which influenced what would have happened in the relevant counterfactual situation. The same pattern of results holds for situations in which the relevant counterfactual was manipulated by turning the teleport on (clip 6) or off (clip 2). Similarly, when ball B went through the gate, causal judgments differed depending on whether the brick would have been in the way (compare clips 14 and 18), or whether the teleport was switched off (clip 13) or on (clip 17).

While the general pattern of judgments was very similar for both question order conditions, there were also some differences. Participants' judgments between conditions differed most strongly for clips 12, 17, and 18. Participants who made counterfactual judgments first gave very low causal ratings for these clips, whereas participants who made causal judgments first gave somewhat higher ratings.

Discussion

The results of Experiment 1 show that causal judgments about particular events are fundamentally linked to counterfactuals. The more certain participants were that the outcome would have been different without ball A, the higher their causal rating. How close the outcome actually was (e.g. whether ball B went right through the middle of the gate, or just barely in) didn't affect participants' judgments. However, participants' judgments differed strongly between clips that were matched in terms of what actually happened but were different in terms of what would have happened if the candidate cause had been removed from the scene. By assuming that people use their intuitive understanding of physics to simulate what would have happened, and that their mental simulations of the underlying physics are somewhat noisy, we were able to capture people's counterfactual judgments very accurately. Even though we don't normally encounter teleports in our everyday lives, participants had no trouble simulating what would have happened in counterfactual situations that included the operation of the teleport.

The results also revealed an interesting order effect. There was a closer correspondence between counterfactual and causal judgments for participants who answered the counterfactual questions in the first block. What explains this order effect? One possibility is that, depending on the question order, participants had different subjective degrees of belief that ball A was a whether-cause by the time they made their causal judgments. Participants who answered the counterfactual question first, had more experience with the physical setup before they made their causal judgments. Remember also that we provided participants with

feedback in the counterfactual block. That is, we showed them what would have happened if ball A had not been present in the scene. Participants who made causal judgments before being asked to make counterfactual judgments may have been more uncertain about what would have happened at this point in the experiment.

Another possibility is that participants who were explicitly asked to make counterfactual judgments in the first block, consequently focused on whether-causation when making causal judgments. Participants who made causal judgments first may have focused more on how-causation and thus assigned greater causality even in situations in which ball A's presence made no difference to whether ball B went through the gate. The fact that the judgments between conditions only differed for causation and not for prevention suggests that how-causation may play a more important role for cause than prevention judgments.

Experiment 2: Two candidate causes

Experiment 1 looked at situations that featured a single candidate cause, and the results showed that causal judgments were strongly influenced by the extent to which the candidate cause was perceived as a whether-cause of the outcome. However, the clips in this experiment did not manipulate the other aspects of causation that the CSM postulates. Experiment 2 tests the CSM more comprehensively by looking at situations that involve two candidate causes. This setup allows us to tease apart the different aspects of causation and test how they affect participants' causal judgments.

Methods

The clips in Experiment 2 included two candidate causes, ball A and ball B, and one candidate target, ball E. Figure 12 shows diagrams of the 32 different clips that participants viewed in this experiment. Remember that for Experiment 1, we created the different clips by contrasting the closeness of the actual outcome with how close the outcome would have been if the candidate cause had been removed from the scene. This time, we constructed the clips by manipulating whether ball E actually went through the gate or missed the gate, as well as whether ball E would have gone through the gate or would have missed it in the relevant counterfactual situations in which either ball A, or ball B, or both balls had been absent. Given that there are four relevant 'worlds' (the actual world, a world with only ball A, only ball B, or one with neither ball A nor ball B) for which ball E can either go through the gate or miss the gate, there are 16 qualitatively different situations in total. For each type of situation, we created two different clips (see Table A1 for detailed information about each clip).

For example, clip 1 shown in Figure 12 shows a case in which ball E missed the gate. Ball E would also not have gone through the gate if either ball B or ball A (or both) had been removed from the scene. For clip 23, in contrast, ball E went through the gate, it would not have gone through the gate if ball B had been removed from the scene, it would have gone through if ball A had been removed, and it would also have gone through if both ball A and B had been removed. This richer setup with two candidate causes allows us to reconstruct many of the situations that have been discussed in the philosophical literature on causation, such as situations of joint causation (clip 3), overdetermination (clip 15), preemption (clip 16), and double prevention (clip 23). Like in Experiment 1,

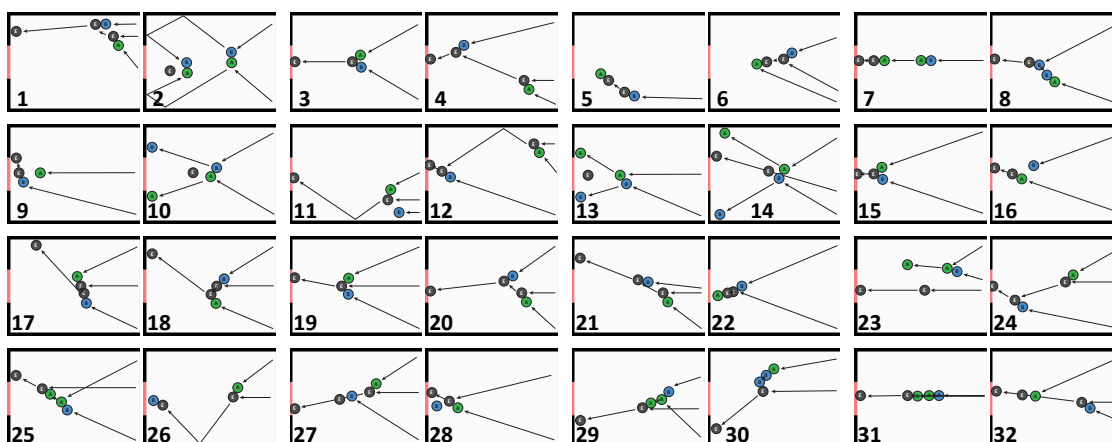


Figure 12. Diagrams of the clips used in Experiment 2. The clips varied whether ball E went through the gate in the actual situation, and what would have happened if either ball A, ball B, or both balls had been removed from the scene. See Table A1 for more information about each clip.

we manipulated between participants whether they were asked to make counterfactual, or causal judgments.

Counterfactual judgments

Determining whether-causation in Experiment 1 required extrapolating where ball B would have gone if ball A had been removed from the scene. However, in this experiment which features three balls, simulating the consequences of removing one ball from the scene is more challenging. Now, the relevant counterfactual simulation may involve collision events between the two remaining balls. For example, if asked whether ball B made a difference to whether or not ball E went through the gate in clip 23 (the double prevention clip, see Figure 12), one needs to simulate whether ball A would have collided with ball E, and whether this collision would have made ball E miss the gate. Here, we look at whether participants are capable of simulating what would have happened in these more complex situations.

Participants. 80 participants ($M = 33$ years, $SD = 10.1$, 34 female) participated in the experiment.

Procedure. Half of the participants made counterfactual judgments about ball A, and the other half about ball B. Participants were instructed that they would see 32 different video clips in total. The order of the clips was randomized. Participants viewed each clip twice before answering the question: “Would ball E have gone through the gate if ball A [ball B] had not been present?”. Participants indicated their response on a slider whose endpoints were labeled “definitely no” and “definitely yes”. The midpoint was labeled “unsure”. After having answered the question, participants received feedback by viewing the same clip again whereby either ball A or ball B was turned into a ‘ghost ball’ that did not collide with the other balls and stopped moving at the point at which it would have first collided. On average, it took participants 18.08 minutes ($SD = 4.63$) to complete the experiment.

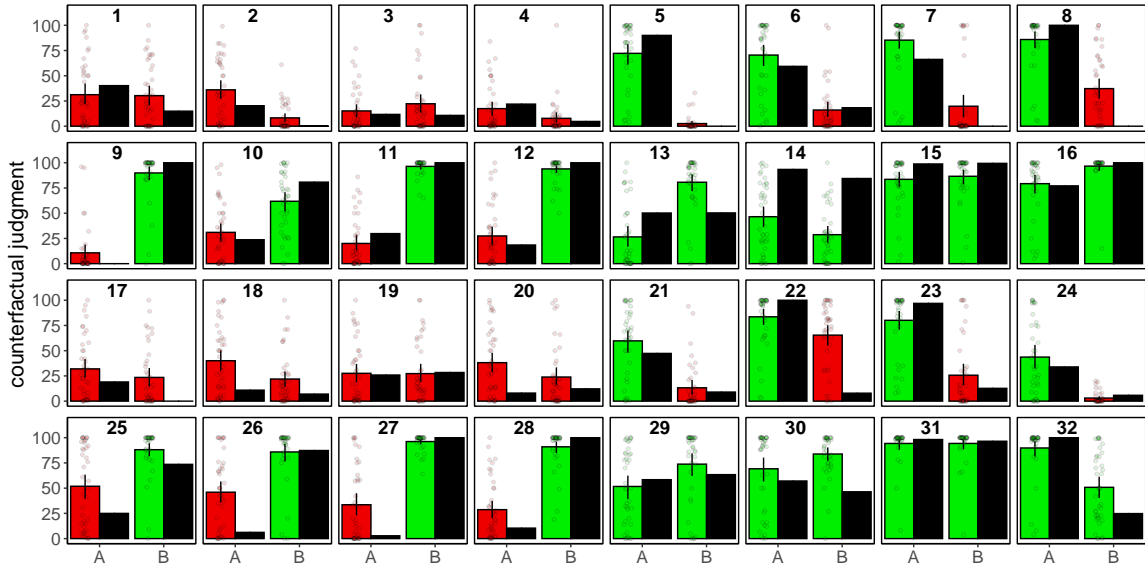


Figure 13. Mean counterfactual judgments (colored bars) together with the model predictions (black bars) of the best-fitting approximate simulation model. For each clip, the two colored bars indicate participants’ belief that ball E would have gone through the gate if ball A had been removed from the scene (left bar) or if ball B had been removed (right bar). For example, in clip 11, participants believed that ball E would not have gone through the gate if ball A had been removed, but would have gone through the gate if ball B had been removed. *Note:* Red bars indicate cases in which ball E would have missed the gate. Green bars indicate cases in which ball E would have gone through. Small points are individual judgments (jittered along the x-axis to increase visibility). Error bars are bootstrapped 95% confidence intervals.

Results. Figure 13 shows participants’ mean counterfactual judgments together with the predictions of the approximate simulation model. To predict participants’ counterfactual judgments, the approximate simulation model first removes the candidate ball from the scene, and then adds noise to the directions of the remaining balls’ velocity vectors at whatever time point their trajectory in the counterfactual situation differs from their actual trajectory. Overall, the approximate simulation model fits participants’ counterfactual judgments well with $r = .86$ and $\text{RMSE} = 19.84$ using a noise parameter of $\theta = 2^\circ$. A deterministic physics model (i.e. $\theta = 0^\circ$) does worse with $r = .83$, $\text{RMSE} = 30.28$.

Discussion. Overall, participants’ counterfactual judgments were again well described by the approximate simulation model even for these more complex clips in which the relevant counterfactual simulation involved considering how multiple balls would have moved. Note that the degree of noise that was required to best account for participants’ judgments ($\theta = 2^\circ$) was higher than in Experiment 1 ($\theta = 0.9^\circ$). This is likely due to the fact that simulating the relevant counterfactuals now may involve computing collisions between the two remaining balls. For example, in clip 14, participants have to simulate how ball A (or ball B) would have collided with ball E and what the outcome of that collision would have been.

There are a few cases for which participants’ counterfactual judgments deviated from

the model’s predictions. In situations in which ball E would not have gone through the gate, the model was often more confident than participants were (e.g., the judgments for ball A in clips 25–28). When evaluating whether ball E would have gone through the gate if ball B had been removed from clip 22, participants have to gauge whether ball E would have managed to pass by ball A. Participants considered it likely that ball E would have gone in. However, it would have required a significant amount of perturbation (in the correct direction) for ball E to pass by ball A, and the model thus predicts that participants’ judgment in this case should be low. Similarly in clip 13, the model predicts that participants should be equally uncertain about whether ball E would have gone through the gate if either ball A or ball B had been removed. However, participants considered it more likely that ball E would have gone through the gate without ball B than without ball A. Whereas ball A is headed for the stationary ball E along a close to horizontal path, ball B comes in at an angle. Ball E would have gone through the gate if either of the balls had been present but participants have greater difficulty extrapolating ball B’s path compared to ball A’s. Overall, while there are a few cases for which model predictions and participants’ judgments come apart, the model again does a fine job at capturing participants’ beliefs about what would have happened.

Causal responsibility judgments

Based on participants’ judgments in the counterfactual condition, we can determine the extent to which each candidate cause qualifies as a whether-cause of E’s going through the gate (see Figure 6). To be able to apply the full CSM, we still need to determine whether each cause was a difference-maker, a how-cause, a sufficient-cause, and a robust-cause of the outcome (see Figure 5). We will first discuss in detail how the CSM was implemented, and then look at how well it accounted for participants’ judgments.

Model predictions. To generate predictions from the CSM, the model first records the clip’s events of interest (e.g. when collisions happen, where and when ball E goes through the gate or misses the gate, etc.). The model then simulates for each ball A and ball B, whether they were difference-makers, and then tests for how-causation, whether-causation, sufficient-causation, and robust-causation. We ran 1000 simulations for each ball and aspect of causation (except for how-causation which is binary and therefore only requires a single simulation to run).

To assess difference-making, the CSM removes the candidate cause from the scene, and then simulates whether the outcome event would have been any different from what actually happened. We construe the outcome event Δe finely and record the exact position at which ball E went through the gate (or missed the gate), as well as the time at which the outcome happened. To capture people’s uncertainty about what would have happened in the relevant counterfactual situation, the CSM applies noise to the direction of motion of the alternative cause at the time at which its trajectory in the counterfactual situation would have been different from the trajectory it actually had (see Figure 1 on how noise is applied in the model). The model then checks for each of the simulations that it runs, whether the candidate cause was a difference-maker of the outcome by comparing the actual outcome with the counterfactual simulation.

For how-causation, the CSM applies a very small perturbation to the ball’s initial position and then checks whether the outcome event Δe would have been different from what it was in the actual clip. If so, the candidate ball qualifies as a how-cause of the

outcome. Rather than yielding a continuous measure, each ball either is a how-cause or isn't (see Table A1). Testing for how-causation captures whether there was a direct transfer of force between a candidate cause and ball E, or whether there was an indirect transfer of force, as in a causal chain where ball B collides with ball A, and ball A subsequently collides with ball E (see (Figure 12, clip 7).

To test for whether-causation, the model removes the candidate cause. To capture people's uncertainty in what would have happened, the model applies noise to the remaining balls from the time at which their movement in the counterfactual world differs from what it was like in the actual world. For example, in clip 24, ball A collides with ball E (at $t = 129$) which subsequently collides with ball B (at $t = 218$) and then goes through the gate. To simulate what would have happened if ball A had been removed, the model applies noise to ball E's motion at $t = 129$, and then to ball B's motion at $t = 218$. The model records whether the outcome (broadly construed) would have been different from what actually happened. For example, if ball E went through the gate in the actual situation, but it would have missed in a counterfactual simulation in which ball A had been removed from the scene, then ball A qualifies as a whether-cause of E's going through in this simulation.

To test for sufficient-causation, the model first removes all alternative causes from the scene. For example, when the model tests whether ball A was a sufficient-cause of ball E's going through the gate, ball B is first removed from the scene. The model then simulates what would have happened in a situation in which only the candidate cause (ball A), and the target (ball E) had been present and records whether ball E would still have gone through the gate or whether it would have missed the gate. Again, uncertainty is introduced into the model by applying noise to the remaining balls' movements at whatever time their movement in the counterfactual situation differs from their movement in the actual situation. Finally, the model considers a situation in which the candidate cause had also been removed from the scene and records the outcome in that situation. Ball A qualifies as a sufficient-cause of E's going through the gate if ball E would still have gone through the gate in a situation in which ball B had been removed *and* ball E would have missed the gate if both ball A and ball B had been removed.

The test for robust-causation is analogous to sufficient-causation. However, instead of removing the alternative cause from the scene, the model applies a small perturbation to the alternative cause's initial position. It then checks what the outcome would have been in this counterfactual situation, and in a situation in which the alternative cause had been perturbed and the candidate cause had been removed. Noise to the ball's motions is applied just like for sufficient-causation and whether-causation. A ball qualifies as a robust-cause if the outcome in the counterfactual situation in which the alternative cause was perturbed would have been the same as actually happened, but the outcome would have been different if the alternative cause was perturbed and the candidate cause was removed.

We will compare three different versions of the CSM. Each version of the model considers whether a cause was a difference-maker, and tests for the different aspects of causation (see Figure 5). The CSM_W only considers whether-causation, the CSM_{WH} considers both whether-causation and how-causation, and the CSM_{WHS} considers whether-causation, how-causation, and sufficient-causation. We did not consider a model that includes robust-causation as an additional predictor because for our selection of clips, whether-causation and robust-causation were very highly correlated (see Table 2). Nevertheless, we believe

that robustness is an important aspect of causation and we will return to this aspect in the General Discussion.

The CSM predicts that the extent to which participants judge a candidate cause as having been causally responsible for the outcome increases the more the different aspects of causation apply. To fit the model to participants’ judgments, we ran Bayesian mixed effects models that infer the weights on each aspect of causation that best account for the data (see Equation 6). The CSM has one free parameter θ for the degree of noise that is applied to each balls’ movement in the counterfactual simulations, one parameter α for a global intercept, and one parameter each β for the different weights on the aspects of causation. So, depending on how many aspects are considered, the model has between two and four free parameters. We determine θ based on participants’ judgments in the counterfactual condition, and then use the same value of θ for all aspects of causation that introduce noise (i.e. difference-making, whether-causation, and sufficient-causation).

Participants. 41 participants ($M = 34$ years, $SD = 10.5$, 21 female) took part in this experiment.

Procedure. The order in which the 32 clips were presented was randomized. Participants viewed each clip three times before answering the question: “To what extent were A and B responsible for E [not] going through the gate?”. The question was adapted based on the outcome of the clip. Participants indicated their responses on two separate sliders – one for each ball – that were presented on the same screen. The endpoints of the sliders were labeled “not at all” and “very much”.⁸ Each slider could be set independently (e.g. giving a low or high rating for both balls). On average, it took participants 21.19 minutes ($SD = 4.96$) to complete the experiment.

Results. Figure 14 shows participants’ mean causal responsibility judgments for each of the two balls in the 32 clips together with the predictions of the CSM_{WHS} that considers whether-causation, how-causation, as well as sufficient-causation. Figure 15 shows how well each of the three versions of the CSM account for the data overall.⁹ We compared

Table 2

Pearson’s r correlation between the different pairs of predictors in the model. Note: The pairwise correlations shown here are before the difference-making aspect is multiplied with the other predictors as specified in Equation 6.

	difference	whether	how	sufficient
whether	.50			
how	.79	.27		
sufficient	.21	.10	.36	
robust	.43	.93	.24	.20

⁸In a replication study, we asked participants to what extent they agreed with the statement: “Ball A/B caused ball E to go through the gate” or “Ball A/B prevented ball E from going through the gate” depending on the outcome. Participants’ causal judgments in this study were highly correlated with participants’ responsibility judgments reported here, suggesting that both question framings elicit very similar judgments. See the online materials for more details <https://github.com/cic1-stanford/csm>.

⁹Table 3 shows the parameters for the different models and Table A1 in the Appendix shows each model’s predictions for the all the different clips as well as the values of the different aspects of causation.

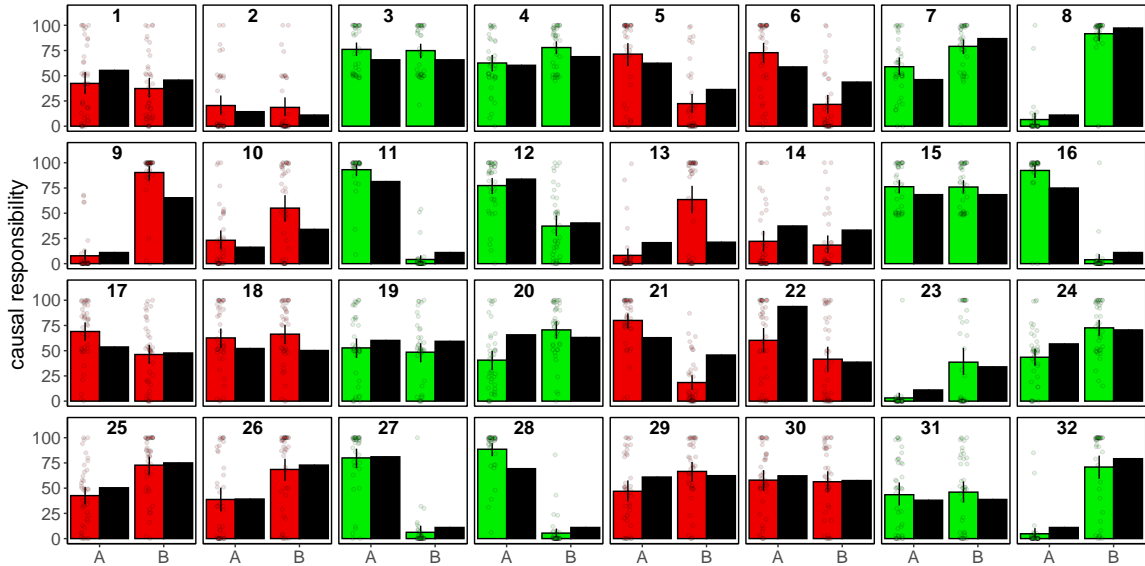


Figure 14. Mean causal responsibility (red = negative outcome, green = positive outcome) and model predictions by CSM_{WHS} (black bars). Note: Small points are individual judgments (jittered along the x-axis to increase visibility). Error bars are bootstrapped 95% confidence intervals.

the models using approximate leave-one-out cross-validation as model selection criterion. According to this criterion, the CSM_{WHS} performs better than the CSM_{WH} (difference in expected log predictive density (elpd) = -107.0 , with a standard error of 14.7) and the CSM_{W} (difference in elpd = -383.5 , standard error = 30.1).¹⁰

To get a better sense for why the different aspects of causation are required for explaining participants’ judgments, we will focus on the five clips that we discussed in the introduction. Figure 16 shows these clips together with participants’ mean judgments as well as the predictions of the different versions of the CSM.

CSM_{W} which only considers whether-causation as a predictor has trouble accounting for several aspects of the data. For example, in the causal chain, it underpredicts judgments about ball A. Even though participants are relatively certain that ball E would still have gone through the gate if ball A had been removed, they still give it a high causal rating. The model also underpredicts participants’ judgments for overdetermination and preemption. In both of these cases, ball E would still have gone through the gate if either of the balls had been removed from the scene. The only way for CSM_{W} to assign any responsibility here is by assuming a high baseline rating (i.e. the model’s intercept; see Table 3). However, this leads the model to make exaggerated predictions for ball A in the double prevention case, and ball B in the preemption case.

CSM_{WH} combines both whether-causation and how-causation to predict participants’ judgments. Taking into account how-causation resolves many of CSM_{W} ’s problems. For

¹⁰As a general rule, a model is considered superior when the magnitude of the negative difference in expected log predictive density is greater than twice the standard error of that difference (for details, see Bürkner & Vuorre, 2019; Vehtari et al., 2017).

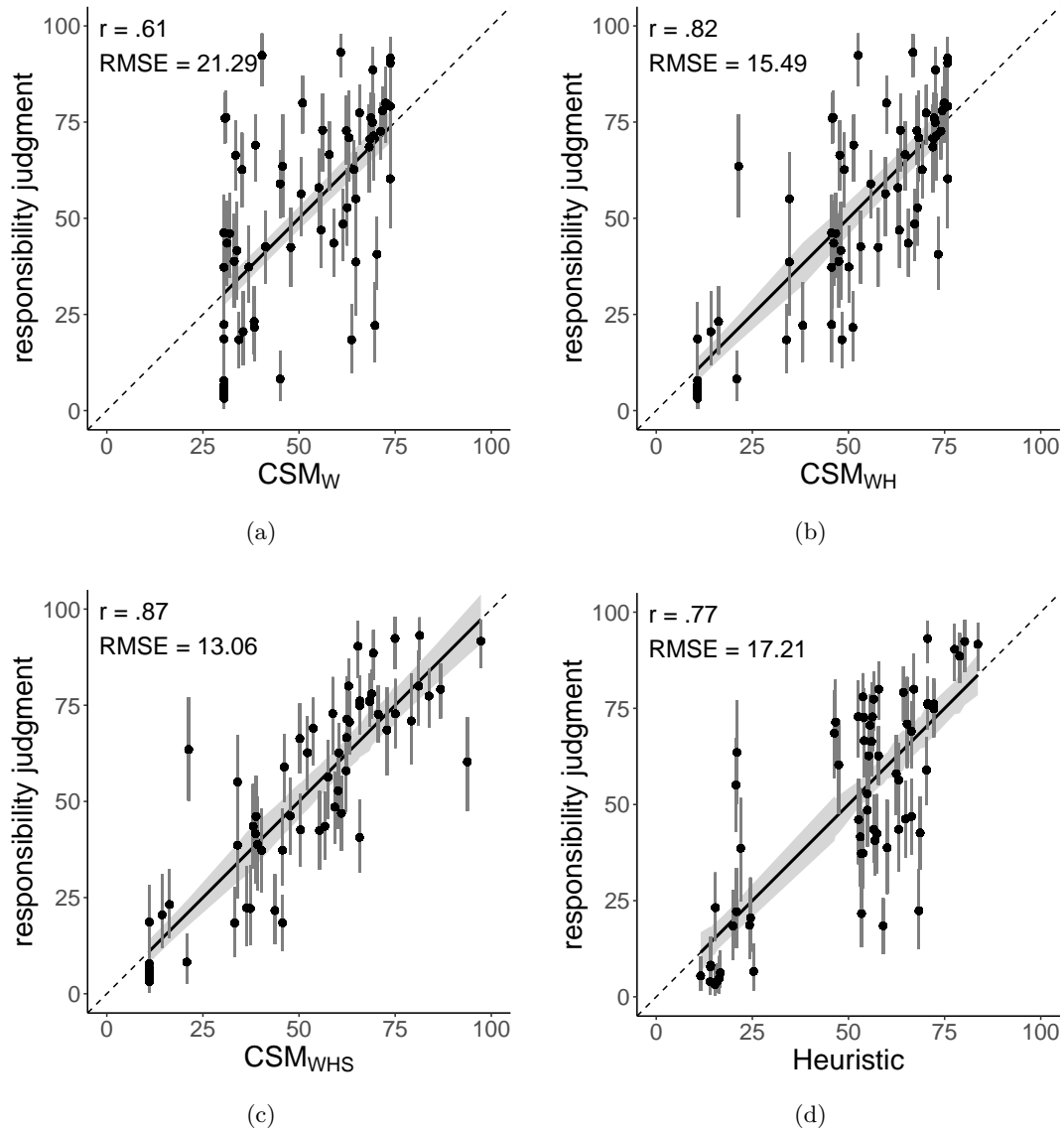


Figure 15. Scatter plot between different versions of the counterfactual simulation model (CSM) as well as the heuristic model (x-axis), and mean causal responsibility judgments (y-axis). *Note:* The gray ribbons show the 95% highest density interval of the regression line. The error bars are bootstrapped 95% confidence intervals. r = Pearson moment correlation, RMSE = root mean squared error, W = whether-causation, H = how-causation, S = sufficient-causation.

example, CSM_{WH} better captures participants' judgments for the causal chain. By taking into account that ball A was a how-cause, and ball B was both a how-cause and a whether-cause, CSM_{WH} explains this pattern of results without assuming a high baseline rating (see Table 3). CSM_{WH} also explains ball B's low rating in the double prevention case. Ball B was only a whether-cause but not a how-cause. Considering how-causation also

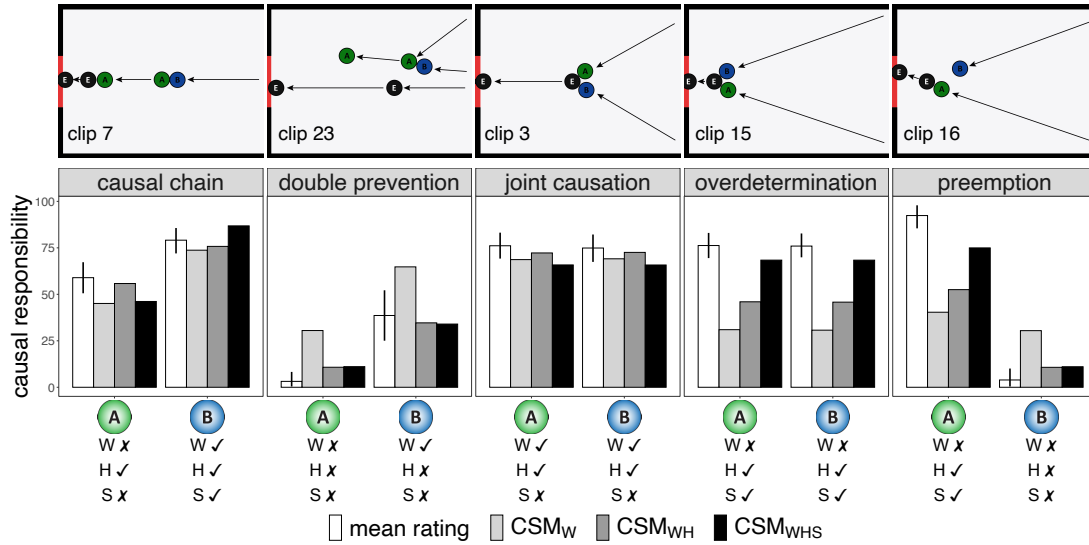


Figure 16. Participants’ mean causal responsibility judgments (white bars) for the same selection of clips as shown in Table 1. The gray bars show the predictions of different versions of the counterfactual simulation models. *Note:* Error bars are bootstrapped 95% confidence intervals. W = whether-causation, H = how-causation, S = sufficient-causation. ✓ = true, ✗ = false.

helps somewhat with overdetermination and preemption. While CSM_{WH} does better than CSM_W, there are still some patterns that the model struggles with. In particular, CSM_{WH} predicts higher ratings for joint causation (where both balls are whether-causes) compared to overdetermination (where neither ball is a whether-cause).

By taking sufficient-causation into account, CSM_{WHS} better explains participants’ judgments in the joint causation and overdetermination case. In joint causation, both balls

Table 3

Causal responsibility judgments as modeled by CSM: Estimates of the mean, standard error, and 95% highest density intervals of the different predictors in the Bayesian mixed effects models.

model	term	estimate	std.error	lower 95% HDI	upper 95% HDI
CSM _w	intercept	30.46	1.62	27.27	33.64
CSM _w	whether	43.28	2.20	38.98	47.62
CSM _{wh}	intercept	10.73	1.54	7.70	13.81
CSM _{wh}	whether	30.17	2.69	24.96	35.39
CSM _{wh}	how	34.88	2.61	29.73	39.95
CSM _{whs}	intercept	11.04	1.55	8.03	14.10
CSM _{whs}	whether	28.96	2.72	23.62	34.40
CSM _{whs}	how	25.32	3.02	19.34	31.33
CSM _{whs}	sufficient	31.97	2.98	26.14	37.97

are whether-causes but neither are sufficient for making ball E go through the gate. The opposite holds for the overdetermination. Here, neither ball is a whether-cause but both balls are sufficient-causes. By postulating that participants care about both necessity and sufficiency, CSM_{WHS} correctly predicts that the ratings should be similarly high in both cases.

Even though CSM_{WHS} provides a very good account of participants' judgments overall, there are still some cases with which the model struggles. For instance, in the preemption case, participants give a very high rating to ball A but the model's prediction is lower. Ball A doesn't qualify as a whether-cause in this case but participants generally care about whether-causation. We will return to the problem of preemption in the General Discussion.

Heuristic model. So far we have focused our analysis on the CSM. We have seen that the simple CSM_W does a very good job of explaining participants' causal judgments in Experiment 1. The more complex CSM_{WHS} accounts well for participants' judgments in Experiment 2. However, we don't know yet whether a model that just relies on information about what actually happened might also explain participants' judgments well in these more complex cases. It is possible, in principle, that participants use different strategies for making causal judgments in simple versus complex cases.

White (2014) suggested that people's causal judgments are sensitive to a number of clues. The core hypothesis is that people's original source of causal knowledge stems from experiencing oneself as an agent acting on objects in the world (White, 2009, 2012a). From these experiences, more abstract features of causal interactions are derived that then serve as clues for identifying causal relationships. Events will be seen as causal to the extent that they resemble these clues.

The clues to causality are heuristic in that they provide useful guides under conditions of uncertainty but do not necessarily identify causal relationships correctly. Generally, the model predicts that the more clues apply to a particular event, the more likely this event will be judged as causal. In an experiment, White (2014) tested the heuristic by showing participants a list of descriptions such as "Two moving cars collide and rebound.", or "A ball rolls down a slope." and asked whether or not they believed that the event was a causal relation. As predicted, participants' cause judgments were highly correlated with the number of causal clues that applied to different events.

White (2014) listed 15 clues for identifying causal relationships. Some clues do not apply to our experiments, such as whether human action was involved, or whether there was corresponding evidence from multiple modalities. Other clues do not discriminate between the different clips that we showed to participants (e.g. the duration of causal interaction which is always instantaneous in our case). Table 4 shows the ten clues that applied to our clips and briefly describes how we implemented them.

We defined a Heuristic model that uses a linear combination of these features to predict participants' judgments. We fit the model to participants' judgments using a Bayesian mixed effects regression model with random intercepts for participants. We imposed constraints on the predictors by specifying priors reflecting the assumption that the predictors positively affect causal judgments (we reverse-coded the predictor "initial movement of E"). Table 5 shows the estimates for the different predictors in the model.

The results show that the most important predictor of participants' causal responsibility judgments was whether force was transferred from cause to effect. Note that this

Table 4

Features of the heuristic model of causal judgment. The “direction” column indicates whether the variable is a positive or negative predictor of causal judgments. The “implementation” column explains how each variable was implemented.

#	variable name	direction	implementation
1	prior movement	+	Dummy variable for whether A/B was initially moving
2	initial speed	+	Initial speed of A/B
3	contact with E	+	Dummy variable for whether A/B contacted E
4	change of E’s speed	+	Difference between E’s speed before and after collision with A/B
5	change of E’s movement direction	+	Difference between E’s direction of motion before and after collision with A/B (measured in angular rotation)
6	change of other objects’ speed	+	Sum of the differences in other objects’ speeds before and after collisions with A/B
7	change of other objects’ movement direction	+	Sum of differences between other objects’ directions of motion before and after collisions with A/B
8	transfer of force	+	Dummy variable for whether A/B transferred force to E
9	initial movement of E	–	Dummy variable for whether E was initially moving
10	exclusive contact with E	+	Dummy variable for whether A/B was the only ball contacting E

predictor is identical to how-cause in the CSM. The other predictors had a substantially smaller effect on the model predictions. Despite its greater number of predictors, the Heuristic model doesn’t capture participants’ judgments as well as the CSM_{WH5} does (see Figure 15).

Individual differences. So far we have only evaluated the models on aggregated judgments. To what extent do individual participants’ judgments differ, and can the CSM explain such differences? The CSM has two ways of accounting for individual differences. First, participants may differ in their beliefs about the different aspects of causation for any given case. Second, participants may differ in how much the different aspects of causation affect their causal judgments. Here, we explore the latter option.

To analyze the extent to which participants differed in their causal responsibility judgments, we ran the CSM on each individual participant’s responses. For each participant, we looked at how well their judgments were explained by the three different versions of the CSM. To fit the models, we z-scored individual participants’ judgments, and restricted the priors over the regression weights to be positive. We used approximate leave-one-out crossvalidation to determine which model best captures each participant’s judgments. We

found that out of the 41 participants' judgments, 39 were best explained by CSM_{WHS} , and 2 by CSM_{WH} . Across individual participants, CSM_W yielded a median correlation of $r = .43$ [5% quantile = .22, 95% quantile = .60], compared to CSM_{WH} with $r = .60$ [.37, .78], and CSM_{WHS} with $r = .64$ [.40, .79].¹¹

To better appreciate the individual differences, Figure 17 shows individual participants' judgments for the selection of five clips which we already discussed above. Each solid line shows an individual participant's judgments for ball A and ball B in this clip. The dashed lines indicate the mean responses across participants. To cluster participants, we first fit the CSM_{WHS} to participants' responses as a Bayesian mixed effects model with random intercepts and random slopes. We then used k-means clustering to assign participants into clusters based on their random effects, finding that a solution with $k = 2$ clusters led to a stable assignment. This clustering analysis revealed that one group of participants cared mostly about how-causation, whereas the other group of participants cared more about whether-causation and sufficient-causation.

For many situations, there was considerable variance in participants' judgments. For example, in the "causal chain" (clip 7), many participants gave a higher rating to ball B than ball A. However, there was also a group of participants who judged both balls equally, either rating both balls around 50 or 100. There was also one participant who gave a much greater rating to ball A than ball B, most likely having confused the two balls during the judgment phase. We can make sense of this inter-individual variation by assuming that participants differ in how much the aspects of how-causation and whether-causation affect their judgment. Participants who mostly cared about how-causation gave similar ratings to both balls whereas participants who cared more about whether-causation rated ball B

Table 5

Causal responsibility judgments as modeled by the Heuristic: Estimates of the mean, standard error, and 95% highest density intervals for the posterior distribution of the different predictors in the Bayesian mixed effects model. The model included random intercepts for participants. Each predictor was z-scored. Note: We put half-Gaussian priors with a standard deviation of 10 and a lower bound of 0 on the predictors to reflect the assumption that the different predictors positively affect causal responsibility judgments.

term	estimate	std.error	lower 95% HDI	upper 95% HDI
intercept	49.73	1.48	46.87	52.70
prior movement	0.22	0.21	0.00	0.79
initial speed	2.08	0.84	0.45	3.73
contact with E	0.38	0.36	0.01	1.35
change of E's speed	0.12	0.12	0.00	0.46
change of E's direction	1.06	0.73	0.06	2.76
change of other objects' speed	2.19	0.95	0.42	4.09
change of other objects' movement direction	3.99	0.90	2.21	5.69
transfer of force	15.59	0.80	13.98	17.14
initial movement of E	0.18	0.18	0.00	0.65
exclusive contact with E	4.38	0.71	3.00	5.79

¹¹Table A2 in the Appendix shows how well each of the models account for each individual participant's judgments.

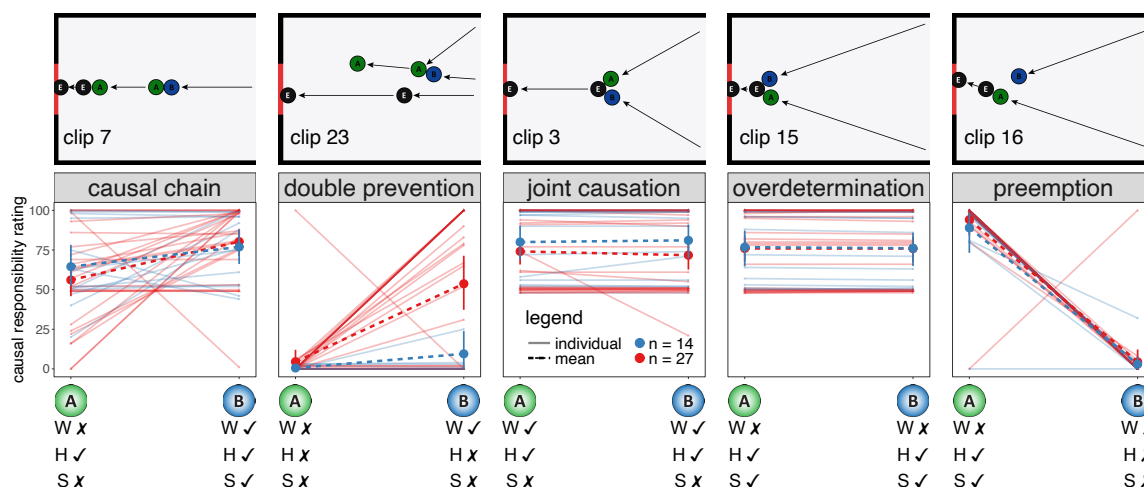


Figure 17. Plots illustrating individual differences in participants’ causal responsibility judgments for a selection of the clips shown in Experiment 2. *Note:* Thin solid lines indicate individual participants’ ratings for balls A and B. Thick dashed lines show cluster means. Red = participants who emphasize how-causation ($N = 16$), blue = participants who emphasize whether-causation and sufficient-causation ($N = 25$). Error bars are bootstrapped 95% confidence intervals. W = whether-causation, H = how-causation, S = sufficient-causation. ✓ = true, ✗ = false.

higher than ball A.

In “double prevention” (clip 23), both groups of participants viewed ball A as not having been responsible at all for ball E’s going through the gate. However, the groups differed in how much causal responsibility they assigned to ball B. Whereas one group rated ball B close to zero (with a few exceptions), another group gave ball B a high rating. Again, the CSM helps us make sense of this variation. There are two ways to explain the variance in ratings to ball B. First, participants may differ in how they weigh how-causation and whether-causation. Participants for whom how-causation is critical gave a low rating to ball B. Participants who care about whether-causation rated ball B highly. Second, some participants may have been unsure whether ball B was actually a whether-cause of ball E’s going through the gate. The counterfactual judgments for this clip show that some participants believed that ball E would have gone through the gate even if ball B had been removed from the scene (see Figure 13).

The pattern of individual differences is less clear in “joint causation” (clip 3) and “overdetermination” (clip 15). Most participants either gave a 50 rating, or a 100 rating to each ball for these clips. The CSM can in principle account for this pattern. Participants who mostly care about how-causation should give high ratings in both cases. If a participant only cared about whether-causation, she should give a higher rating in “joint causation” than in “overdetermination”. However, we know that most participants care about sufficient-causation also, and at least on the aggregate level, both factors appear to influence participants’ judgments to a similar extent (see Table 3).

The clips discussed so far suggest that there was considerable variance in how causal judgments were reached. The results for “preemption” (clip 16) show that this wasn’t always

the case. The pattern of results here was strikingly consistent: almost all participants gave a close-to-maximal rating to ball A and a zero rating to ball B. As discussed above, the CSM struggles with explaining this pattern since ball A was not a whether-cause of ball E's going through in this case, but whether-causation is an important factor for how people make causal judgments across a variety of situations.

Discussion

Experiment 2 provided a challenging test for the CSM. By expanding our setup to include two candidate causes, we were able to reconstruct many of the situations that have troubled traditional counterfactual theories of causation. The results show that the CSM was up to the challenge. The model accounted for participants' causal responsibility judgments to a high degree of quantitative accuracy. Whereas in Experiment 1, the CSM only needed to consider whether-causation in order to explain participants' judgments, Experiment 2 demonstrated that participants are sensitive to different aspects of causation when making causal judgments. Overall, a model that considered whether-causation, how-causation, and sufficient-causation struck the best balance between model complexity and fit.

Even though the CSM explained participants' judgments well in Experiment 1 it was still possible in principle that people would use different strategies to make causal judgments for more complex interactions that feature more than a single candidate cause. To find out, we compared the CSM to a heuristic model that predicts causal judgments based on a number of features (e.g. each ball's velocity, force transfer, contacts, etc.) that capture what actually happened in the clip, and that have been argued to influence how people make causal judgments about specific events (see White, 2014). This model performed worse than the CSM_{WHS} despite having more free parameters to fit the data. Overall, this shows that a model which aims to capture causal judgments without relying on counterfactual contrasts fails to adequately explain participants' responses.

We also looked at individual participants' judgments and saw that there was considerable variation. According to the CSM, this variation may arise from systematic differences in how much participants take the different aspects of causation into account when making causal judgments. Currently, we can only speculate what drives these individual differences. One possibility is that participants interpreted the question differently. We asked participants "To what extent were A and B responsible for E going through the gate?". Participants who focused on the "how" may have interpreted this question as meaning "going through the gate *in the way that it did*", whereas "whether"-participants may have interpreted the question to mean going through versus *not* going through the gate. Another possibility is that people operate with different intuitions about what makes for a good cause. If this was the case, we might see individual tendencies of focusing on "how" versus "whether" to show up in other domains as well. Another source of individual differences may come from variation in participants' beliefs about what would have happened in the relevant counterfactual situations. It is possible that participants put more emphasis on how-causation in situations in which they find it difficult to compute whether-causation (or sufficient-causation). This would suggest a context-specific weighting of the different aspects of causation where aspects that an observer is more certain about influence causal judgments more strongly.

General Discussion

This paper introduced a computational model of how people make causal judgments about physical events: the *counterfactual simulation model* (CSM). The CSM accurately predicts participants' causal judgments across a wide range of dynamic physical scenes featuring single and multiple causes, as well as familiar objects (like balls, walls, and blocks) and unfamiliar ones (like the teleport). The central claim of the model is that causal judgments are intimately linked to counterfactual simulations (see Gerstenberg, Peterson, et al., 2017). While this claim is not new (see Hart & Honoré, 1959/1985; Kahneman & Tversky, 1982; Lewis, 1973; Lipe, 1991; Mackie, 1974), previous models that have linked causal judgments to counterfactuals were troubled by situations in which people see an event as causal even though the outcome would still have happened if that event hadn't come about. Some have argued that people's intuitions in such situations of causal overdetermination demonstrate that causal judgments are dissociated from counterfactuals (Mandel, 2003; Mandel & Lehman, 1996). Others have suggested that people operate with several fundamentally different notions of causation (Hall, 2004; Lombrozo, 2010). Here, we have shown how an enriched counterfactual model that considers multiple aspects of causation handles situations that have troubled previous accounts, and that it does so in a way that bridges previous conceptions of causation (Strevens, 2013).

The CSM builds on interventionist theories of causation (e.g. Pearl, 2000; Woodward, 2003) and assumes that people have a causal mental model of the world that supports simulating the consequences of different counterfactual interventions. This generative mental model dictates the causal processes that govern how the world unfolds. Traditionally, interventionist accounts have expressed domain knowledge with fairly abstract representations, such as structural equations (Halpern, 2016; Halpern & Pearl, 2005). In this paper, we have proposed a different strategy. Instead of representing people's causal models with binary variables and abstract functions that relate these variables, we assume that people have a rich mental representation of the scene that captures the dynamics of the actual situation (Woodward, 2011a). For the dynamic collision events considered here, people's domain knowledge can be expressed as an intuitive model of physics that is similar to modern game engines that generate physically realistic simulation environments (see Battaglia et al., 2013; Smith & Vul, 2013; Ullman et al., 2017).

Using this mental game-engine, people can bring to mind different counterfactuals by simulating how the situation would have unfolded if objects had been removed from the scene, or if something about these objects had been changed (see also Chater & Oaksford, 2013; Gerstenberg & Tenenbaum, 2017; Goodman et al., 2015). These different counterfactual simulations reveal various aspects of causation that express the different ways in which a cause affected the outcome. We get a notion of *whether-causation* by simulating whether the outcome would have been qualitatively different if the candidate cause had been removed from the scene. Most counterfactual theories have focused exclusively on this aspect of causation. However, by casting people's causal representation of the scene in terms of a physics engine that supports counterfactual interventions, we derive additional aspects of causation that previous counterfactual theories neglected. We get a notion of *how-causation* by considering whether a small perturbation to the candidate cause would have made a difference to the outcome event (finely construed). Intuitively, *how-causation*

captures whether causal events are connected in a more direct way – this is the notion of causation that process theories focus on. By considering counterfactual interventions on the alternative causes in the scene, we define the aspects of sufficient-causation and robust-causation. A cause is sufficient if it would still have brought about the outcome even if alternative causes had been removed from the scene, and robust to the extent that it would still have brought the outcome if the alternative causes had been somewhat perturbed.

The results of one preliminary study and two experiments support the CSM. In the preliminary study, participants’ cause and prevention judgments increased the more certain they were that the outcome would have been different if the candidate cause had been removed. This study featured a broad range of clips where what actually happened differed in each clip. This means that it might in principle be possible to explain people’s causal judgments just in terms of what actually happened and without reference to counterfactual contrasts.

To provide a more rigorous test for the role of counterfactual simulation in causal judgments, Experiment 1 featured pairs of clips in which what actually happened was identical, but what would have happened in the relevant counterfactual situation was different. For example, depending on the position of a block, the ball would have either gone through the gate or it would have missed. The CSM correctly predicted that participants’ judgments between these pairs of clips would differ as a function of what would have happened, even though what actually happened was held constant.

In Experiment 2, we looked at situations that featured two candidate causes. In this setting, we were able to reconstruct many of the problem cases that have been discussed in the philosophical literature on causation, such as situations of double prevention, overdetermination, and preemption. The results of this experiment show that people’s causal judgments are sensitive to different aspects of causation, and that people differ in how much they rely on each aspect when making causal judgments. A heuristic model that uses a variety of features which capture what actually happened (see White, 2014) didn’t account for participants’ judgments as well. Counterfactual contrasts appear to be necessary for explaining participants’ causal judgments even in more complex settings.

Table 6 shows a qualitative comparison of the different accounts. The CSM is the only model that yields quantitative predictions. The theories differ in the extent to which they consider counterfactuals. Whereas the original account of the *force dynamics model* tried to do away with counterfactuals completely (Wolff, 2007), a more recent formulation of the model introduced the counterfactual notion of a “virtual force” as one way to handle causation by omission (Wolff et al., 2010). Structural equation models are inadequate for representing a more detailed understanding of the mechanistic processes that underlie people’s causal judgments about physical events (Jensen, 2019). The feature-based model is unable to adequately handle situations that involve multiple causes. Finally, whereas structural equation models have been usefully employed in a variety of domains, the other models presented here still have to prove whether they generalize beyond the physical domain.

Limitations and open challenges

In the remainder, we will discuss limitations of the CSM as well as open challenges. We will talk about (1) the process by which people reach causal judgments, (2) the implications of treating objects versus events as the locus of counterfactual interventions, (3) how the

Table 6

Qualitative comparison of different models of causal judgment.

	force dynamics model	structural equation model	feature-based model	counterfactual simulation model
quantitative predictions	no	no	no	yes
considers counterfactuals	somewhat	yes	no	yes
captures processes	yes	no	yes	yes
handles multiple causes	yes	yes	no	yes
generalizes beyond physical causation	somewhat	yes	no	somewhat

CSM’s causal aspects may help elucidate the semantics of causal verbs like “caused” and “enabled”, (4) how normative expectations may influence what counterfactuals come to mind, (5) what to do about the problem of preemption, (6) the relationship between causal judgments and causal perception, (7) what function causal judgments play, and finally (8) how the CSM applies outside the physical domain.

The process of making causal judgments. The CSM postulates that people make causal judgments by mentally simulating what would have happened in relevant counterfactual situations and comparing the simulated outcome to what actually happened. Recently, we have shown that this account not only accurately predicts participants’ causal judgments but that it also captures the cognitive process by which people reach these judgments. In Gerstenberg, Peterson, et al. (2017), we tracked participants’ eye-movements while they were watching clips similar to those of Experiment 1 (but without a brick or teleport, see Figure 2 for examples). Between participants, we varied what question participants were asked to answer about the clip. In the “outcome” condition, participants had to judge whether ball B went right through the middle of the gate (if it went through), or whether it completely missed the gate (if it missed). The “counterfactual” and a “causal” conditions were similar to the ones used here in Experiment 1. We found a striking difference in participants’ eye-movements between the “outcome” condition and the other two conditions. Whereas in the “outcome” condition, participants mostly just looked at ball B and sometimes tried to predict where ball B would go next, participants’ eye-movements in the other conditions strongly suggest that participants simulated where ball B would have gone if ball A hadn’t been present in the scene. The extent to which participants engaged in these counterfactual looks was modulated by how uncertain the counterfactual outcome was, showing more looks in situations in which it was difficult to tell what would have happened. Even though participants in the “causal” condition weren’t told anything about counterfactuals in the instructions, they spontaneously engaged in counterfactual simulation in the service of reaching a causal judgment.

Causal relata: Objects vs. events. The CSM treats objects as causes of events. Ball A caused ball B to go through the gate (rather than the collision between ball A and ball B). We share this focus on objects (or agents) as the target of analysis with Wolff’s (2007) *force dynamics model*. Focusing on objects (rather than events) as the unit of analysis also resonates well with White’s (2009) proposal that the experience of acting upon objects in the world shapes our understanding of causality (see also Mayrhofer & Waldmann, 2014).

The structural equation account of causal judgment (Halpern, 2016; Halpern & Pearl, 2005) models counterfactual inferences by considering interventions on variables representing events (e.g. what if the collision had not happened). Instead, the CSM considers

interventions on objects (e.g. what if the ball had not been present in the scene). A key advantage of defining interventions on objects is that they lead to well-defined counterfactuals in our setting. While there are many ways to bring about a counterfactual situation in which a collision event didn't happen (e.g. stopping one of the balls just before the collision, changing a ball's angle, turning a ball into a ghost ball, ...), a situation in which ball A had been removed is well-specified. Game engines may provide a good starting point for exploring what kinds of counterfactuals spontaneously come to people's minds when imagining how things could have turned out differently (Ullman et al., 2017, see also Phillips & Cushman, 2017). Some interventions in a game engine are easier to realize than others. We can add or remove objects, make them go faster or slower, make them heavier or lighter, change their elasticity, friction, etc. However, we cannot directly intervene on events such as the collision between two balls for example.

By defining interventions on objects rather than events, we can also make sense of related concepts such as physical support (see Battaglia et al., 2013; Hamrick et al., 2016). In Gerstenberg, Zhou, et al. (2017) we showed participants blocks stacked on a table, and they were asked to say how responsible one of the blocks was for the others staying on the table. Another group of participants saw the same scenes and was asked to say how many of the other blocks would fall off the table if the block was removed. The results revealed a very close correspondence between responsibility judgments and hypothetical predictions about what would happen if the block was removed. A block was seen as more responsible as the proportion of other blocks that were predicted to fall increased. Alternative models that predicted participants' responsibility judgments based on scene features, such as the tower's height, or the position of the to-be-removed block, did not do as well.

These results suggest deep similarities between judgments of causation, and judgments of physical support. Both cognitive processes can be understood as involving an intervention on the generative model of the scene, and a subsequent mental simulation of how the world would have played out. What it means for one object to support another is to prevent it from falling (or cause it to be stable). Note that in the case of physical support the scene is static and there are no events.¹² By considering interventions on objects, both judgments about dynamic causation and static support are handled in a unified way.

The language of causation. People use many different causal verbs to describe what happened (Abelson & Kanouse, 1966; Brown & Fish, 1983; Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017; McDermott, 1995; Rudolph & Forsterling, 1997; Solstad & Bott, 2017; Van Valin & Wilkins, 1996), such as “caused”, “enabled”, or “helped” (Lombard, 1990; Mackie, 1992). Different accounts have been developed to capture the semantics of “caused” versus “enabled”, some within the mental model theory tradition (Goldvarg & Johnson-Laird, 2001), and others using structural equations (Sloman, Barbey, & Hotaling, 2009). Wolff's (2007) force dynamics model differentiates not only “caused” from “enabled” but also other causal expressions such as “prevented” and “despite” based on the force configurations that these verbs map onto. In the force dynamics model, “enabled” is distinguished from “caused” by way of the patient's tendency and the alignment between agent and patient force.

The CSM suggests a new perspective on the semantics of different causal verbs (see

¹²At least no events in the intuitive sense describing changes of state. In philosophy, events are often construed more broadly such that “a block lying on the table” could count as an event (see e.g. Lewis, 1987).

Beller, Bennett, & Gerstenberg, submitted; Gerstenberg & Tenenbaum, 2016, 2017). Instead of mapping different causal verbs onto the space of force configurations, the CSM suggests a mapping onto its multidimensional space of causal aspects. Different causal expressions occupy different regions within that space. For example, just considering the dimensions of whether-causation and how-causation, we can draw distinctions between “caused”, “enabled”, and “affected”. “Caused” applies best when both aspects of causation are true. “Enabled” is specifically sensitive to whether-causation but does not require how-causation. If ball A knocks ball B out of the way such that ball E can go through the gate, ball A enabled (but didn’t cause) ball E to go through the gate (see Freitas et al., 2017; McCawley, 1978; Sloman et al., 2009). “Affected” is the weakest out of the three terms. In order to have affected the outcome, it suffices to have been a how-cause. For example, if ball E is already headed toward the gate and ball A collides with ball E to speed it up, then ball A affected ball E’s going through the gate, but didn’t cause or enable it. Beller et al. (submitted) show that a model which defines a semantics of these different causal expressions using the CSM’s aspects accurately captures participants’ judgments about which expression best explains what happened across a range of video clips similar to the ones we use here.

Normative expectations. The CSM currently doesn’t incorporate normative expectations. Probabilities only enter the model by representing uncertainty about what would have happened in the relevant counterfactual situations. These probabilities are dictated by the observer’s understanding of the situation. For example, an observer who does not know how teleports work will reach different causal judgments from an observer who does.

Research has shown that people’s causal judgments are not only affected by their subjective degree of belief about what would have happened in relevant counterfactual situations, but also by their expectations about what will happen in the actual situation (e.g. Johnson & Rips, 2015; Knobe & Fraser, 2008; Sytsma, Livengood, & Rose, 2012). In general, people have a tendency to select abnormal over normal events as the cause of an outcome (Gerstenberg, Halpern, & Tenenbaum, 2015; Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard et al., 2017; Kominsky et al., 2015). While the preference for abnormal causes has long been noted as an empirical phenomenon (Hart & Honoré, 1959/1985; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009), there are now also a number of accounts that quantify how normality considerations affect causal judgments (Halpern & Hitchcock, 2015; Icard et al., 2017; Kominsky et al., 2015; Morris et al., 2018).

To explain participants’ causal judgments in the experiments discussed here, we did not need to incorporate normative expectations about what will happen. One reason for this might be that we employed balanced experimental designs. In both experiments reported in this paper, a candidate ball was equally likely to serve as a cause of the outcome, prevent it from happening, or make no difference. However, the billiard ball paradigm provides an excellent test bed for exploring effects of expectations on participants’ causal judgments. For example, if participants learn that one ball generally prevents another from going through the gate, while another ball generally tends to make it go through, one could expect asymmetric judgments in a situation in which both balls jointly caused another one to go through the gate. It will also be interesting to see whether the effects of normality that have thus far been demonstrated in vignette studies will generalize to the domain con-

sidered here (see Gerstenberg & Icard, 2019; Kirfel et al., in prep, for evidence that this may be the case).

The problem of preemption. Even though the CSM captures people’s causal judgments across a wide range of situations, it still struggles with some cases. In the preemption scenario (Collins, 2000; McDermott, 1995), ball A knocks ball E through the gate shortly before ball B would have done the same (see Figure 3e). The model cannot account for the fact that the preempting cause (ball A) receives a causal rating that is close to ceiling. The reason the CSM fails here is that the preempting cause was not a whether-cause of the outcome, and whether-causation generally matters for people’s causal judgments. That the presence of the preempted cause makes no difference to how the preempting cause is evaluated is often taken as evidence in favor of process theories over counterfactual theories of causation (Paul, 1998a, 1998b; Paul & Hall, 2013). However, the results of the experiments reported in this paper make it clear that counterfactual contrasts are necessary for explaining causal judgments. So, what can we do about preemption?

One possible solution is that how important the different aspects of causation are varies between situations. However, this only pushes the problem up a level in that one would now need a meta-level theory that predicts in what situations the different aspects are important. Alternatively, it could be that people’s causal representation of the situation differs from what actually happened. At first sight, the following solution is tempting: people construct a reduced causal representation of the situation that only features those aspects of the situation that qualified as a cause of the outcome (using the difference-making test in Equation 1). Since the preempted cause doesn’t qualify as a cause, it’s as if it was never present in the scene. In such a situation the preempting cause *would* have also been a whether-cause of the outcome, and hence the model would predict a maximum rating.

While tempting, this simple solution doesn’t work. In Experiment 1 we saw that participants’ causal judgments differed strongly as a function of where the brick was placed in the scene. For example, in neither clip 14 nor clip 18 does the brick make any difference to what actually happens (see Figure 9). However, the causal ratings differed dramatically between the two cases. Participants say that ball A cause ball B to go through the gate in clip 14, but not in clip 18. If we were to simply remove the brick from the scene (because it didn’t make any difference in the actual situation), we wouldn’t be able to capture the key difference between these two scenes anymore.

That being said, there is an important difference between the case of causation involving the brick, and the preemption case. In the *brick case*, the brick is a *potential preventer* – it would have prevented the ball from going through if the collision between the balls hadn’t taken place. In the *preemption case*, the other object is a *potential cause*: the other ball would have also caused ball E to go through the gate – just a moment later. So one possibility is the following: in situations of redundant causation that feature several potential causes but where only one cause directly affected the outcome, the other potential causes that didn’t affect the outcome are removed from the causal representation of the scene.

Dealing with preemption is tricky. While the CSM provides some tools for tackling these cases, it doesn’t have all the answers and more work is required (Collins, 2000; McDermott, 1995).

Causal judgments vs. causal perception. Sometimes we perceive causation directly (Blakemore et al., 2001; Hubbard, 2012a, 2012b; Michotte, 1946/1963; Rips, 2011;

Saxe & Carey, 2006; Schlottmann, 2000; Scholl & Tremoulet, 2000; White, 2012b). When we see two billiard balls colliding, we have a direct impression that one caused the other to move. This causal inference is immediate and doesn't appear to involve any counterfactual simulation. In fact, it has been shown that there is retinotopic adaptation to causal perception events (Kominsky & Scholl, 2020; Rolfs, Dambacher, & Cavanagh, 2013) suggesting that certain causal events are encoded very early in visual processing. What is the relationship between causal perception and causal judgments?

One intriguing possibility is that both causal perception and causal judgments rely on inference but that the inference required for causal perception is just much faster (and easier) than that for causal judgments (see Bechlivanidis, Schlottmann, & Lagnado, 2019). Even though it doesn't feel like we're engaging in counterfactual simulation when perceiving that one ball launched another, it's worth noting that in these cases, computing the relevant counterfactual simulation is extremely simple and fast. We know that if ball A hadn't struck the stationary ball B, then ball B would have just stayed like it was. So, to judge whether ball A caused ball B's movement, an observer only needs to run a counterfactual simulation for a single step (and has already seen what this counterfactual would look like). If, in contrast, an observer wants to judge whether ball A caused ball B to go through a gate like in our experiments, the relevant counterfactual simulation now spans across a longer temporal duration, and involves generating a situation that wasn't already seen before.

Prior research has shown that perceiving and judging causation sometimes come apart (Levillain & Bonatti, 2011; Rips, 2011; Schlottmann, 1999; Schlottmann & Shanks, 1992,?; Thorstad & Wolff, 2016; Wolff & Shepard, 2013). Wolff and Shepard (2013) propose a dual process approach which generates fast feelings of causation that can later be overcome by knowledge about possible causal mechanisms. They report an experiment in which participants view an animation of a person hitting a fire hydrant, and the lights of a town in the background turning off. Participants were asked to judge whether they "felt for a moment" that the blackout was caused by the person hitting the fire hydrant, and whether they "ultimately concluded" so. When the blackout happened at the same time as the hitting event, almost all participants answered "yes" to the feel-question, and the majority answered "no" to the conclude-question.

This dual process view is consistent with the sequential way in which the CSM considers the different aspects of causation. We might rely on our causal perception system to first detect which objects were "a cause" of the outcome, and then later take into account the different aspects of causation, like whether-causation and sufficient-causation, to ultimately determine the extent to which each candidate object was actually "the cause" of the outcome. Such a model could explain dissociations between judgments of causal perception and causal necessity (Schlottmann & Shanks, 1992). While judgments of causal perception only require to establish a causal connection, judgments of necessity are linked to a consideration of whether-causation which requires a more involved counterfactual simulation.

The function of causal judgments. A good theory of causation is one that helps us to adequately characterize how people learn, reason, plan, and act upon the world (Woodward, 2014). It is easy to justify why we should have an accurate causal model of how the world works. Such a model serves as a guide for action. To reach our goals, we need to consider what the likely consequences of different actions on the world would be. This process requires a generative model of how the world works, and a way of simulating

the consequences of hypothetical interventions. The CSM assumes that people already have access to a generative model of the domain. Recent work has looked into how it may be possible to learn such a causal model from observing and interacting with the world (Baradel, Neverova, Mille, Mori, & Wolf, 2019; Battaglia et al., 2018; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Ullman et al., 2018; Yi* et al., 2020). If a person already has an accurate model of the world, what role, if any, does the ability to make causal judgments play?

Causal judgments form the foundation for assigning moral and legal responsibility (Hart & Honoré, 1959/1985; Lagnado & Gerstenberg, 2017; Moore, 2009; Stapleton, 2008; Wright, 1985). To hold someone responsible, one has to establish that the person's action played a causal role in bringing about the outcome (Alicke, 2000; Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Chockler & Halpern, 2004; Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Lagnado, 2010, 2012; Halpern, 2016; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Lagnado, Gerstenberg, & Zultan, 2013; Malle, Guglielmo, & Monroe, 2014; Niemi, Hartshorne, Gerstenberg, & Young, 2016; Shaver, 1985; Weiner, 1995; Zultan, Gerstenberg, & Lagnado, 2012), and holding others responsible is important for regulating interpersonal relationships (Forsyth & Kelley, 1994; Lewis, 1948; Rai & Fiske, 2011).

Causal judgments are also intimately linked to explanations (Hilton, 2007; Lewis, 1986a; Lombrozo, 2006, 2010, 2012), and there is evidence that the act of generating explanations benefits learning (Lombrozo, 2016; Lombrozo & Carey, 2006). Many explanations are of a causal nature. For example, we postulate reasons as causes when explaining why a person acted how they did (Buss, 1978; Malle, 1999). When answering *why* something happened, we pick out those events that made a difference to the outcome. The causal judgments of one person can serve as valuable learning input to the other person (Hilton, 1990; Kirfel et al., in prep).

Recently, it has been proposed that causal judgments further play the role of highlighting those aspects that not only made a difference in the actual situation, but that are also likely to continue to make a difference in other situations, too (Danks, 2013; Hitchcock, 2012; Lombrozo, 2010; Nagel & Stephan, 2016). A causal judgment now, may help to pinpoint a useful place for intervention in the future (Bramley, Gerstenberg, & Tenenbaum, 2016; Bramley, Mayrhofer, Gerstenberg, & Lagnado, 2017; Gerstenberg & Icard, 2019; Girotto, Legrenzi, & Rizzo, 1991; Meder, Gerstenberg, Hagmayer, & Waldmann, 2010).

Forming causal representations of a situation also allows us to communicate efficiently what happened. Hearing that “Tom broke the vase” versus “Tom allowed the vase to break”, results in different mental models of what happened (e.g. Freitas et al., 2017; Solstad & Bott, 2017; Van Valin & Wilkins, 1996).

Going beyond physics. In this paper, we applied the CSM to modeling causal judgments about physical events. However, causality doesn't only happen between billiard balls. It happens between people, markets, countries, etc. (Hagmayer & Osman, 2012). What does the CSM have to say about causal judgments outside the physical domain?

On the most general level, the CSM maintains that causal judgments about specific events are best understood as counterfactual contrasts over generative models. We believe that this insight applies broadly. Probabilistic programs are a powerful tool for building generative models (see, e.g. Ellis et al., 2020; Goodman et al., 2015; Lake, Salakhutdinov, &

Tenenbaum, 2015). The physics engine that the CSM uses is such a probabilistic program, but the approach is much more general. For example, probabilistic programs have been used to explain people’s inferences about the interactions of social agents as well (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Evans, Stuhlmüller, Salvatier, & Filan, 2017; Gerstenberg & Tenenbaum, 2017; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, submitted; Stuhlmüller & Goodman, 2014; Ullman et al., 2009; Zhi-Xuan, Mann, Silver, Tenenbaum, & Mansinghka, 2020). Inference in probabilistic programs often takes the form of running simulations. By injecting noise into these simulations, probabilistic programs naturally handle counterfactual inferences in a probabilistic way (Perov et al., 2020; Tavares, Koppel, Zhang, & Solar-Lezama, 2018).

We further believe that the specific aspects of causation that the CSM postulates may be useful for illuminating what factors influence people’s causal judgments beyond the physical domain. For example, people who emphasize how-causation might differentiate more strongly between acts of omission versus commission (Gerstenberg & Stephan, 2020; Livengood & Machery, 2007; McGrath, 2005; Spranca, Minsk, & Baron, 1991; Stephan, Willemsen, & Gerstenberg, 2017), or pay particular attention to the role of force in harmful events when judging the morality of actions (Cushman & Young, 2011; Greene et al., 2009; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Iliev, Sachdeva, & Medin, 2012; Mikhail, 2007). The aspect of robust-causation is related to the role that intentions play in how actions are evaluated (Heider, 1958; Lombrozo, 2010; Woodward, 2006). We generally blame others more for negative outcomes that resulted from intended versus accidental actions (Grinfeld et al., 2020; Heider, 1958; Kleiman-Weiner et al., 2015; Malle, 2004; Malle et al., 2014; Uhlmann, Pizarro, & Diermeier, 2015). Whereas accidents are sensitive to the particular situational circumstances, an agent will adapt their actions as the situation changes to bring about an intended outcome. Intentions make the causal relationship between actions and outcomes *robust* (Heider, 1958; Lombrozo, 2010).

How the CSM’s counterfactual operators are implemented in domains outside of intuitive physics needs to be worked out. For example, if Sarah helps John with studying for his exam, then she is a how-cause of the exam outcome. She may not have made a difference to whether John passed, but she made a difference to how well he did (see McDermott, 1995). Of course, we would not consider a *change()* here as a perturbation in physical space, but rather a perturbation to Sarah’s actions (e.g. her helping more or less).

When we think about people as causes there are also other kinds of counterfactuals that come to mind. For example, it might not only matter whether a person did something, and how they did it, but also how someone else would have acted in the same situation (see Falk & Szech, 2013). Imagining what the reasonable person would have done is a common procedure in the law (Green, 1967; Lagnado & Gerstenberg, 2017), and may also play a role in how people assign responsibility (Fincham & Jaspars, 1983; Gerstenberg, Ullman, et al., 2018).

Considering how to apply the CSM outside of physics highlights an important limitation. If the generative model of the domain is inaccurate, then the counterfactual simulations that the CSM generates will be inaccurate too, and so will be the causal judgments it makes. Some domains may involve much more fundamental levels of uncertainty, in which our knowledge is incomplete and erroneous, and where the workings of the system

are influenced by many unknown factors. For example, lacking an accurate causal model of the financial system, the CSM can't say whether the fall of Lehman Brothers caused the financial crisis. It is also conceivable that an observer has an accurate causal model of the domain but that it's just not clear what the relevant counterfactual contrast should be. The CSM doesn't solve the problem of selecting which counterfactual contrasts are relevant for a particular situation (Kominsky & Phillips, 2019). Given the probabilistic nature of the CSM's different aspects of causation, the model can in principle account for multiple sources of uncertainty. However, it remains to be seen in practice how well the model extends beyond the physical.

Conclusion

How do people make causal judgments about physical events? This paper presents a novel theory: the *counterfactual simulation model* (CSM). The CSM makes three key assumptions: (1) causal judgments are about difference-making, (2) difference-making for particular events is best expressed in terms of counterfactual contrasts over causal models, and (3) there are multiple aspects of causation which correspond to different ways of making a difference to the outcome that jointly determine people's causal judgments about physical events.

As a case study, we applied the CSM to explain people's causal judgments about dynamic collision events. The results revealed that people's judgments are influenced by different aspects of causation, such as whether the candidate cause was necessary and sufficient for the outcome to occur, as well as whether it affected how the outcome came about. By modeling these aspects in terms of counterfactual contrasts, the CSM accurately captures participants' judgments in a wide variety of physical scenes involving single and multiple causes.

Some important challenges remain. Future versions of the CSM will need to incorporate the role of normative expectations, more adequately handle situations of preemption, and capture people's judgments in domains that go beyond physical causation.

Acknowledgments

We thank all the whether-causes without whom this paper wouldn't have happened, and the how-causes who helped to improve it. In particular, we thank Ari Beller, Beth Levin, Christopher Hitchcock, Christos Bechlivanidis, Dan Lassiter, Felipe De Brigard, Fiery Cushman, Jonas Nagel, Jonathan Kominsky, Jonathan Phillips, Jonathan Schaffer, Joseph Halpern, Joshua Hartshorne, Joshua Knobe, Julian De Freitas, Kevin Smith, Laurie Paul, Liang Zhou, Max Kleiman-Weiner, Members of CICL, Members of CoCoSci, Members of Lagnado Lab, Members of SHAME writing group, Ned Hall, Neil Bramley, Nori Jacobi, the participants of the TaCitS conference (2017), the participants of the London Judgment and Decision Making seminar, the participants of the Metaphysics Ranch workshop (2015), the participants of the Modality workshop at Yale (2014), Paul Bello, Pascale Willemsen, Peter Battaglia, Phillip Wolff, Ralf Mayrhofer, Richard Holten, Shaun Nichols, Henrik Singmann, Simon Stephan, Steven Sloman, Susan Carey, Thomas Icard, and Tomer Ullman. This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216, as well as by ONR awards N00014-09-1-0124, N00014-13-1-0333, and N00014-13-1-0788.

Data from the preliminary study, Experiment 1 and 2 have appeared in the Proceedings of the Annual Conference of the Cognitive Science Society in Gerstenberg et al. (2012), Gerstenberg, Goodman, Lagnado, and Tenenbaum (2014), and Gerstenberg, Goodman, Lagnado, and Tenenbaum (2015), respectively.

References

- Abelson, R. P., & Kanouse, D. E. (1966). Subjective acceptance of verbal generalizations. In S. Feldman (Ed.), *Cognitive consistency: Motivational antecedents and behavioral consequences* (pp. 171–197). New York: Academic Press.
- Ahn, W.-K., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 199–225). Cambridge, MA: Cambridge University Press.
- Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.
- 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.
- Allen, K., Jara-Ettinger, J., Gerstenberg, T., Kleiman-Weiner, M., & Tenenbaum, J. B. (2015). Go fishing! responsibility judgments when cooperation breaks down. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 84–89). Austin, TX: Cognitive Science Society.
- Aronson, J. L. (1971). On the grammar of ‘cause’. *Synthese*, *22*(3), 414–430.
- Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, mar). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. Retrieved from <https://doi.org/10.1038%2Fs41562-017-0064-017-0064> doi: 10.1038/s41562-017-0064
- Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.
- Baradel, F., Neverova, N., Mille, J., Mori, G., & Wolf, C. (2019). Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*.
- Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., ... Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks*.
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.
- Bear, D., Fan, C., Mrowca, D., Li, Y., Alter, S., Nayebi, A., ... others (2020). Learning physical graph representations from visual scenes. *Advances in Neural Information Processing Systems*, *33*.
- Bechlivanidis, C., Schlottmann, A., & Lagnado, D. A. (2019, April). Causation without realism. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0000602
- Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The oxford handbook of causation*. Oxford University Press, USA.
- Beller, A., Bennett, E., & Gerstenberg, T. (submitted). The language of causation.
- Blakemore, S. J., Fonlupt, P., Pachot-Clouard, M., Darmon, C., Boyer, P., Meltzoff, A. N., ... Decety, J. (2001). How the brain perceives causality: an event-related fmri study. *NeuroReport*, *12*(17), 3741–3746.
- Bramley, N., Gerstenberg, T., & Tenenbaum, J. B. (2016). Natural science: Active learning in dynamic physical microworlds. In A. Papafragou, D. Grodner, D. Mirman, &

- J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2567–2572). Austin, TX: Cognitive Science Society.
- Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–1910.
- Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2019). Intervening in time. In S. Kleinberg (Ed.), *Time and causality across the sciences* (pp. 86–115). Cambridge University Press.
- Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.
- Bramley, N. R., Mayrhofer, R., Gerstenberg, T., & Lagnado, D. A. (2017). Causal learning from interventions and dynamics in continuous time. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 150–155). Austin, TX: Cognitive Science Society.
- Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*(3), 237–273.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01
- Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *25*.
- Buss, A. R. (1978). Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology*, *36*(11), 1311–1321.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).
- Cartwright, N. (1995). False idealisation: A philosophical threat to scientific method. *Philosophical Studies*, *77*(2), 339–352.
- Cartwright, N. (2004). Causation: One word, many things. *Philosophy of Science*, *71*(5), 805–820. Retrieved from <https://doi.org/10.1086%2F426771> doi: 10.1086/426771
- Chang, W. (2009). Connecting counterfactual and physical causation. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1983–1987). Cognitive Science Society, Austin, TX.
- Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*(6), 1171–1191.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.
- Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567.
- Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.
- Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382. Retrieved from <http://dx.doi.org/10.1037/0033-295x.99.2.365> doi: 10.1037/0033-295x.99.2.365
- Cheng, P. W., & Novick, L. R. (2005). Constraints and nonconstraints in causal learning:

- Reply to white (2005) and to luhmann and ahn (2005). *Psychological Review*, 112, 694–706.
- Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22(1), 93–115.
- Collins, J. (2000, Apr). Preemptive prevention. *The Journal of Philosophy*, 97(4), 223. Retrieved from <http://dx.doi.org/10.2307/2678391> doi: 10.2307/2678391
- Collins, J. D., Hall, E. J., & Paul, L. A. (2004). *Causation and counterfactuals*. MIT Press Cambridge, MA.
- Craik, K. J. W. (1943). *The nature of explanation*. Cambridge, UK: Cambridge University Press.
- Croft, W. (1991). *Syntactic categories and grammatical relations: The cognitive organization of information*. University of Chicago Press.
- Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, Mar). Evaluating amazon’s mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, 8(3), e57410. Retrieved from <http://dx.doi.org/10.1371/journal.pone.0057410> doi: 10.1371/journal.pone.0057410
- Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, 35(6), 1052–1075.
- Danks, D. (2013). Functions and cognitive bases for the concept of actual causation. *Erkenntnis*, 78(S1), 111–128. Retrieved from <https://doi.org/10.1007/s10670-013-9439-2> doi: 10.1007/s10670-013-9439-2
- Danks, D. (2017). Singular causation. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 201–215). Oxford University Press.
- De Vreese, L. (2006). Pluralism in the philosophy of causation: desideratum or not? *Philosophica*, 77, 5–13.
- Dowe, P. (2000). *Physical causation*. Cambridge, England: Cambridge University Press.
- Dowe, P. (2001, jun). A counterfactual theory of prevention and “causation” by omission. *Australasian Journal of Philosophy*, 79(2), 216–226. Retrieved from <http://dx.doi.org/10.1080/713659223> doi: 10.1080/713659223
- Downing, C. J., Sternberg, R. J., & Ross, B. H. (1985). Multicausal inference: Evaluation of evidence in causally complex situations. *Journal of Experimental Psychology: General*, 114(2), 239–263.
- Ehring, D. (1986). The transference theory of causation. *Synthese*, 67(2), 249–258.
- Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, 99(1), 3–19.
- Ellis, K., Wong, C., Nye, M., Sable-Meyer, M., Cary, L., Morales, L., . . . Tenenbaum, J. B. (2020). Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*.
- Evans, O., Stuhlmüller, A., Salvatier, J., & Filan, D. (2017). *Modeling Agents with Probabilistic Programs*. <http://agentmodels.org>. (Accessed: 2017-5-22)
- Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, 14(3), 219–250.
- Falk, A., & Szech, N. (2013). Morals and markets. *Science*, 340(6133), 707–711.
- Fincham, F. D., & Jaspars, J. M. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology*, 22(2), 145–161.
- Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016, aug). Functional

- neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*(34), E5072–E5081. Retrieved from <http://dx.doi.org/10.1073/pnas.1610344113> doi: 10.1073/pnas.1610344113
- Forsyth, D. R., & Kelley, K. N. (1994). Attribution in groups estimations of personal contributions to collective endeavors. *Small Group Research*, *25*(3), 367–383.
- Freitas, J. D., DeScioli, P., Nemirow, J., Massenkoff, M., & Pinker, S. (2017). Kill or die: Moral judgment alters linguistic coding of causality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Retrieved from <https://doi.org/10.1037/xlm0000369> doi: 10.1037/xlm0000369
- Gerstenberg, T., & Goodman, N. D. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 788–793). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., & Icard, T. F. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*.
- Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.
- Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review*, *19*(4), 729–736.
- Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from <https://doi.org/10.1177/0956797617713053> doi: 10.1177/0956797617713053
- Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2018). What happened? reconstructing the past from vision and sound. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.
- Gerstenberg, T., & Stephan, S. (2020). A counterfactual simulation model of causation by omission. *PsyArXiv*. Retrieved from <https://psyarxiv.com/wmh4c/>

- Gerstenberg, T., & Tenenbaum, J. B. (2016). Understanding “almost”: Empirical and computational studies of near misses. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2777–2782). Austin, TX: Cognitive Science Society.
- Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmann (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, August). Lucky or clever? From expectations to responsibility judgments. *Cognition*, *177*, 122–141. doi: 10.1016/j.cognition.2018.03.019
- Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 409–414). Austin, TX: Cognitive Science Society.
- Giroto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, *78*(1-3), 111–133.
- Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., . . . Zhang, J. (2010). Actual causation: a stone soup essay. *Synthese*, *175*(2), 169–192.
- Glynn, L. (2017). A proposed probabilistic extension of the Halpern and Pearl definition of actual causation. *British Journal for the Philosophy of Science*, *68*, 1061–1124.
- Godfrey-Smith, P. (2010). Causal pluralism. In H. Beebe, C. Hitchcock, & P. Menzies (Eds.), *Oxford handbook of causation* (pp. 326–337). Oxford University Press.
- 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., Mansinghka, V. K., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Uncertainty in Artificial Intelligence*.
- 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.
- Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110.
- Green, E. (1967). The reasonable man: Legal fiction or psychosocial reality? *Law & Society Review*, *2*, 241–258.
- Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009, jun). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. Retrieved from <http://dx.doi.org/10.1016/j.cognition.2009.02.001> doi: 10.1016/j.cognition.2009.02.001
- Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.
- Griffiths, T. L., & Tenenbaum, J. B. (2007, may). From mere coincidences to meaningful discoveries. *Cognition*, *103*(2), 180–226. Retrieved from <http://dx.doi.org/10.1016/j.cognition.2006.03.004> doi: 10.1016/j.cognition.2006.03.004
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.
- Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal

- responsibility and robust causation. *Frontiers in Psychology*, 11, 1069. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2020.01069> doi: 10.3389/fpsyg.2020.01069
- Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: causal bayes nets as rational models of everyday causal reasoning. *Synthese*, 189(1), 17–28.
- Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals*. MIT Press.
- Halpern, J. Y. (2016). *Actual causality*. MIT Press.
- Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, 66, 413–457.
- 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.
- Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, 157, 61–76.
- Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. New York: Oxford University Press.
- Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, 8(6), 280–285.
- Heider, F. (1958). *The psychology of interpersonal relations*. John Wiley & Sons Inc.
- Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019, September). A counterfactual explanation for the action effect in causal judgment. *Cognition*, 190, 157–164. doi: 10.1016/j.cognition.2019.05.006
- 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.
- Hewstone, M., & Jaspars, J. (1987). Covariation and causal attribution: A logical model of the intuitive analysis of variance. *Journal of Personality and Social Psychology*, 53(4), 663.
- Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs*, 39(4), 632–657.
- Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1), 65–81.
- Hilton, D. J. (2007). Causal explanation: From social perception to knowledge-based attribution. In A. Kruglanski & E. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 232–253). New York: Guilford Press.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93(1), 75–88.
- Hitchcock, C. (1995). Salmon on explanatory relevance. *Philosophy of Science*, 62(2), 304–320.
- Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98(6), 273–299.
- 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.
- Hitchcock, C. R. (1996). The role of contrast in causal and explanatory claims. *Synthese*,

- 107(3), 395–419. Retrieved from <https://doi.org/10.1007%2Fbf00413843> doi: 10.1007/bf00413843
- Hoerl, C., McCormack, T., & Beck, S. (2011). *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*. Oxford University Press.
- Hubbard, T. L. (2012a, Dec). Phenomenal causality ii: Integration and implication. *Axiomathes*, 23(3), 485–524. Retrieved from <http://dx.doi.org/10.1007/s10516-012-9200-5> doi: 10.1007/s10516-012-9200-5
- Hubbard, T. L. (2012b, Nov). Phenomenal causality i: Varieties and variables. *Axiomathes*, 23(1), 1–42. Retrieved from <http://dx.doi.org/10.1007/s10516-012-9198-8> doi: 10.1007/s10516-012-9198-8
- Hume, D. (1748/1975). *An enquiry concerning human understanding*. Oxford University Press.
- Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, 161, 80–93. Retrieved from <https://doi.org/10.1016%2Fj.cognition.2017.01.010> doi: 10.1016/j.cognition.2017.01.010
- Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition*, 40(8), 1387–1401.
- Jackson, F. (1977, may). A causal theory of counterfactuals. *Australasian Journal of Philosophy*, 55(1), 3–21. Retrieved from <http://dx.doi.org/10.1080/00048407712341001> doi: 10.1080/00048407712341001
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, 20(10), 785. Retrieved from <https://doi.org/10.1016%2Fj.tics.2016.08.007> doi: 10.1016/j.tics.2016.08.007
- Jaspars, J., Hewstone, M., & Fincham, F. D. (1983). Attribution theory and research: The state of the art. In J. M. Jaspars, F. D. Fincham, & M. Hewstone (Eds.), *Attribution theory and research: Conceptual, developmental and social dimensions* (pp. 343–369). New York: Academic Press.
- Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, 79(1), 1–17.
- Jensen, D. (2019). Overcoming the poverty of mechanism in causal models. In S. Kleinberg (Ed.), *Time and causality across the sciences*. Cambridge University Press.
- Johnson, S. G., & Rips, L. J. (2015, mar). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, 77, 42–76. Retrieved from <http://dx.doi.org/10.1016/j.cogpsych.2015.01.003> doi: 10.1016/j.cogpsych.2015.01.003
- 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.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, 34(7), 1185–1243.
- Kirfel, L., Icard, T. F., & Gerstenberg, T. (in prep). Inference from explanation.
- Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of

- intention and permissibility in moral decision making. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128). Austin, TX: Cognitive Science Society.
- Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The cognitive science of morality: intuition and diversity* (Vol. 2). The MIT Press.
- Kominsky, J. F., & Phillips, J. (2019, Oct). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, *43*(11). Retrieved from <http://dx.doi.org/10.1111/cogs.12792> doi: 10.1111/cogs.12792
- Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.
- Kominsky, J. F., & Scholl, B. J. (2020, October). Retinotopic adaptation reveals distinct categories of causal perception. *Cognition*, *203*, 104339. Retrieved 2020-07-22, from <https://linkinghub.elsevier.com/retrieve/pii/S001002772030158X> doi: 10.1016/j.cognition.2020.104339
- Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017, oct). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. Retrieved from <https://doi.org/10.1016%2Fj.tics.2017.06.002> doi: 10.1016/j.tics.2017.06.002
- Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.
- Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.
- Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856–876.
- Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451–460.
- Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford University Press.
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015, dec). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. Retrieved from <http://dx.doi.org/10.1126/science.aab3050> doi: 10.1126/science.aab3050
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016, Nov). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. Retrieved from <http://dx.doi.org/10.1017/s0140525x16001837> doi: 10.1017/s0140525x16001837
- Levillain, F., & Bonatti, L. L. (2011). A dissociation between judged causality and imagined locations in simple dynamic scenes. *Psychological science*, *22*(5), 674–681.
- Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.
- Lewis, D. (1986a). Causal explanation. *Philosophical Papers*, *2*, 214–240.
- Lewis, D. (1986b). Postscript C to ‘Causation’: (Insensitive causation). In *Philosophical*

- papers* (Vol. 2). Oxford: Oxford University Press.
- Lewis, D. (1987). Events. In *Philosophical papers* (Vol. II, pp. 241–270). Oxford University Press.
- Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, 97(4), 182–197.
- Lewis, H. D. (1948). Collective responsibility. *Philosophy*, 23(84), 3–18.
- Lipe, M. G. (1991). Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin*, 109(3), 456–471.
- Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, 31(1), 107–127.
- Lombard, L. B. (1990). Causes, enablers, and the counterfactual analysis. *Philosophical Studies*, 59(2), 195–211.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10), 464–470.
- Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61(4), 303–332.
- Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford: Oxford University Press.
- Lombrozo, T. (2016, oct). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, 20(10), 748–759. Retrieved from <http://dx.doi.org/10.1016/j.tics.2016.08.001> doi: 10.1016/j.tics.2016.08.001
- Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99(2), 167–204.
- Machamer, P., Darden, L., & Craver, C. F. (2000, mar). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1. Retrieved from <http://dx.doi.org/10.1086/392759> doi: 10.1086/392759
- Mackie, J. L. (1974). *The cement of the universe*. Oxford: Clarendon Press.
- Mackie, P. (1992). Causing, delaying, and hastening: Do rains cause fires? *Mind*, 101(403), 483–500.
- Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, 3(1), 23–48.
- Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.
- Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014, Apr). A theory of blame. *Psychological Inquiry*, 25(2), 147–186. Retrieved from <http://dx.doi.org/10.1080/1047840x.2014.877340> doi: 10.1080/1047840x.2014.877340
- 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.
- Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology*, 71(3), 450–463.
- Mandel, D. R., & Lehman, D. R. (1998). Integration of contingency information in judgments of cause, covariation, and probability. *Journal of Experimental Psychology: General*, 127(3), 269–258.

- Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.
- Mayrhofer, R., & Waldmann, M. R. (2014, May). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cogn Sci*, n/a-n/a. Retrieved from <http://dx.doi.org/10.1111/cogs.12132> doi: 10.1111/cogs.12132
- McCawley, J. D. (1978). Conversational implicature and the lexicon. *Syntax and semantics*, *9*, 245–259.
- McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science*, *46*, 523–544.
- McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, *123*(1), 125–148.
- 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.
- Menzies, P. (1989, mar). A unified account of causal relations. *Australasian Journal of Philosophy*, *67*(1), 59–83. Retrieved from <http://dx.doi.org/10.1080/00048408912343681> doi: 10.1080/00048408912343681
- Michotte, A. (1946/1963). *The perception of causality*. Basic Books.
- Mikhail, J. (2007, apr). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. Retrieved from <http://dx.doi.org/10.1016/j.tics.2006.12.007> doi: 10.1016/j.tics.2006.12.007
- Moore, M. S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford University Press.
- Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions. *PsyArXiv*. Retrieved from <https://psyarxiv.com/nq53z>
- Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806–811). Austin, TX: Cognitive Science Society.
- Ney, A. (2009). Physical causation and difference-making. *The British Journal for the Philosophy of Science*, *60*(4), 737–764.
- Niemi, L., Hartshorne, J., Gerstenberg, T., & Young, L. (2016). Implicit measurement of motivated causal attribution. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1745–1750). Austin, TX: Cognitive Science Society.
- Paul, L. A. (1998a). Keeping track of the time: Emending the counterfactual analysis of causation. *Analysis*, *58*(3), 191–198.
- Paul, L. A. (1998b, jan). Problems with late preemption. *Analysis*, *58*(1), 48–53. Retrieved from <https://doi.org/10.1093/analys/58.1.48> doi: 10.1093/analys/58.1.48
- Paul, L. A. (2000, apr). Aspect causation. *The Journal of Philosophy*, *97*(4), 235. Retrieved from <https://doi.org/10.2307/2678392> doi: 10.2307/2678392
- Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.
- Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their

- identification. *Synthese*, 121(1-2), 93–149.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.
- Perov, Y., Graham, L., Gourgoulias, K., Richens, J., Lee, C., Baker, A., & Johri, S. (2020). Multiverse: causal reasoning using importance sampling in probabilistic programming. In *Symposium on advances in approximate bayesian inference* (pp. 1–36).
- Phillips, J., & Cushman, F. (2017, apr). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 114(18), 4649–4654. Retrieved from <https://doi.org/10.1073/pnas.1619717114> doi: 10.1073/pnas.1619717114
- Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.
- R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 118(1), 57–75. Retrieved from <http://dx.doi.org/10.1037/a0021867> doi: 10.1037/a0021867
- Rips, L. J. (2011). Causation from perception. *Perspectives on Psychological Science*, 6(1), 77–97.
- Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, 121(1), 133–148.
- Rolf, M., Dambacher, M., & Cavanagh, P. (2013, February). Visual adaptation of the perception of causality. *Current Biology*, 23(3), 250–254. doi: 10.1016/j.cub.2012.12.017
- Rudolph, U., & Forsterling, F. (1997). The psychological causality implicit in verbs: A review. *Psychological Bulletin*, 121(2), 192–218. Retrieved from <http://dx.doi.org/10.1037/0033-2909.121.2.192> doi: 10.1037/0033-2909.121.2.192
- Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press, Princeton NJ.
- Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science*, 61(2), 297–312.
- Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, 120(2), 411–437.
- Saxe, R., & Carey, S. (2006, sep). The perception of causality in infancy. *Acta Psychologica*, 123(1-2), 144–165. Retrieved from <https://doi.org/10.1016/j.actpsy.2006.05.005> doi: 10.1016/j.actpsy.2006.05.005
- Schaffer, J. (2000a, jun). Causation by disconnection. *Philosophy of Science*, 67(2), 285. Retrieved from <http://dx.doi.org/10.1086/392776> doi: 10.1086/392776
- Schaffer, J. (2000b, apr). Trumping preemption. *The Journal of Philosophy*, 97(4), 165. Retrieved from <http://dx.doi.org/10.2307/2678388> doi: 10.2307/2678388
- Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, 114(3), 327–358.
- Schlottmann, A. (1999). Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism. *Developmental psychology*, 35, 303–317.

- Schlottmann, A. (2000). Is perception of causality modular? *Trends in Cognitive Sciences*, 4(12), 441–441.
- Schlottmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *The Quarterly Journal of Experimental Psychology*, 44(2), 321–342.
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8), 299–309.
- Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. Springer-Verlag, New York.
- Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, 47(1), 1–51.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press, USA.
- Sloman, S. A., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, 33(1), 21–50.
- Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (submitted). Probabilistic models of physical reasoning.
- Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, 5(1), 185–199.
- Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1467–1472). Austin, TX: Cognitive Science Society.
- Solstad, T., & Bott, O. (2017). Causality and causal reasoning in natural language. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 619–644). Oxford University Press.
- Sosa, F. A., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (submitted). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology.
- Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27(1), 76–105.
- Stapleton, J. (2008). Choosing what we mean by ‘causation’ in the law. *Missouri Law Review*, 73(2), 433–480.
- Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation: A computational model. *Cognitive Science*, 44(7), e12871.
- Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in inaction: Counterfactual simulation and causation by omission. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1132–1137). Austin, TX: Cognitive Science Society.
- Strevens, M. (2013, jul). Causality reunified. *Erkenntnis*, 78(S2), 299–320. Retrieved from <http://dx.doi.org/10.1007/s10670-013-9514-8> doi: 10.1007/s10670-013-9514-8
- Stuhlmüller, A., & Goodman, N. D. (2014). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28, 80–99.

- Suppes, P. (1970). *A probabilistic theory of causation*. Amsterdam: North-Holland.
- Sytsma, J., Livengood, J., & Rose, D. (2012). Two types of typicality: Rethinking the role of statistical typicality in ordinary causal attributions. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(4), 814–820.
- Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.
- Tavares, Z., Koppel, J., Zhang, X., & Solar-Lezama, A. (2018). *A language for counterfactual generative models*.
- Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 301–322). Oxford: Oxford University Press.
- Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.
- Thomason, R. H. (2014). Formal semantics for causal constructions. *Causation in grammatical structures*, 58–75.
- Thorstad, R., & Wolff, P. (2016). What causal illusions might tell us about the identification of causes. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 919–924). Austin, TX: Cognitive Science Society.
- Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.
- Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012, Oct). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480. Retrieved from <http://dx.doi.org/10.1016/j.cogdev.2012.07.005> doi: 10.1016/j.cogdev.2012.07.005
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, sep). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. Retrieved from <https://doi.org/10.1016%2Fj.tics.2017.05.012> doi: 10.1016/j.tics.2017.05.012
- Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.
- Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in Neural Information Processing Systems* (Vol. 22, pp. 1874–1882).
- Van Valin, R. D., & Wilkins, D. (1996). The case for “effector”: Case roles, agents, and agency revisited. *Grammatical constructions: Their form and meaning*, 289–322.
- Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018, April). Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*. doi: 10.1111/cogs.12605
- Vehtari, A., Gelman, A., & Gabry, J. (2017, September). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi: 10.1007/s11222-016-9696-4
- Walsh, C. R., & Sloman, S. A. (2011). The meaning of cause and prevent: The role of causal mechanism. *Mind & Language*, *26*(1), 21–52.
- Waskan, J. A. (2003). Intrinsic cognitive models. *Cognitive science*, *27*(2), 259–283.
- Waskan, J. A. (2011). Mechanistic explanation at the limit. *Synthese*, *183*(3), 389–408.

- Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York: The Guilford Press.
- Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.
- White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*(3), 580–601.
- White, P. A. (2012a). The experience of force: The role of haptic experience of forces in visual perception of object motion and interactions, mental simulation, and motion-related judgments. *Psychological Bulletin*, *138*(4), 589–615.
- White, P. A. (2012b). Visual impressions of causality: Effects of manipulating the direction of the target object's motion in a collision event. *Visual Cognition*, *20*(2), 121–142.
- White, P. A. (2014). Singular clues to causality and their use in human causal judgment. *Cognitive Science*, *38*(1), 38–75. Retrieved from <http://dx.doi.org/10.1111/cogs.12075> doi: 10.1111/cogs.12075
- Williamson, J. (2006). Causal pluralism versus epistemic causality. *Philosophica*, *77*, 69–96.
- Wolff, P. (2003). Direct causation in the linguistic coding and individuation of causal events. *Cognition*, *88*(1), 1–48.
- 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.
- Wolff, P., & Shepard, J. (2013). Causation, touch, and the perception of force. In *Psychology of learning and motivation* (pp. 167–202). Elsevier BV. Retrieved from <http://dx.doi.org/10.1016/b978-0-12-407237-4.00005-0> doi: 10.1016/b978-0-12-407237-4.00005-0
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, England: Oxford University Press.
- Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.
- Woodward, J. (2011a). Mechanisms revisited. *Synthese*, *183*(3), 409–427.
- Woodward, J. (2011b). Psychological studies of causal and counterfactual reasoning. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology*. Oxford: Oxford University Press.
- Woodward, J. (2014). *A functional account of causation*. Retrieved from <http://philsci-archive.pitt.edu/10978/>
- Woodward, J. (2015, Jul). The problem of variable choice. *Synthese*, *193*(4), 1047–1072. Retrieved from <http://dx.doi.org/10.1007/s11229-015-0810-5> doi: 10.1007/s11229-015-0810-5
- Wright, R. W. (1985). Causation in tort law. *California Law Review*, *73*, 1735–1828.
- Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).
- Yablo, S. (2002). De facto dependence. *The Journal of Philosophy*, *99*(3), 130–148.

- Yi*, K., Gan*, C., Li, Y., Kohli, P., Wu, J., Torralba, A., & Tenenbaum, J. B. (2020). Clevrer: Collision events for video representation and reasoning. In *International conference on learning representations*. Retrieved from <https://openreview.net/forum?id=HkxYzANYDB>
- Zhi-Xuan, T., Mann, J., Silver, T., Tenenbaum, J., & Mansinghka, V. (2020). Online bayesian goal inference for boundedly rational planning agents. *Advances in Neural Information Processing Systems*, 33.
- Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, 125(3), 429–440.

Appendix

Table A1

Information about each clip. **Outcome:** both = both balls are present, only A = only ball A is present, only B = only ball B is present, neither = neither A nor B is present, 1 = E goes through the gate, 0 = E misses the gate (For example, in clip 7, the outcome is positive (i.e. ball E goes through the gate) if both balls are present, and if only ball B is present. Otherwise, the outcome is negative.); **Cause:** different aspects of causation; **Model:** predicted ratings of different versions of the counterfactual simulation model (W = whether-cause, H = how-cause, S = sufficient-cause), as well as the heuristic model. **Rating:** mean causal responsibility judgments.

Clip	Ball	Outcome				difference	Cause				Model				Rating
		both	only A	only B	neither		whether	how	sufficient	robust	CSM _W	CSM _{WH}	CSM _{WHS}	Heuristic	
1	A					100	40	100	23	36	48	58	55	57	42
1	B	0	0	0	0	100	15	100	16	9	37	50	46	54	37
2	A					57	12	0	0	10	35	14	14	25	21
2	B	0	0	0	0	18	0	0	0	0	30	11	11	24	19
3	A					100	88	100	12	76	69	72	66	72	76
3	B	1	0	0	0	100	89	100	11	75	69	73	66	72	75
4	A					100	78	100	4	78	64	69	60	58	63
4	B	1	0	0	0	100	95	100	15	57	72	74	69	54	78
5	A					100	90	100	0	47	69	73	62	47	71
5	B	0	0	1	0	100	0	100	0	0	30	46	36	68	22
6	A					100	59	100	16	35	56	64	59	53	73
6	B	0	0	1	0	100	18	100	6	14	38	51	44	53	22
7	A					100	34	100	0	25	45	56	46	70	59
7	B	1	0	1	0	100	100	100	67	60	74	76	87	64	79
8	A					0	0	0	0	0	30	11	11	25	7
8	B	1	0	1	0	100	100	100	100	100	74	76	97	84	92
9	A					0	0	0	0	0	30	11	11	14	8
9	B	0	1	0	0	100	100	100	0	100	74	76	65	78	90
10	A					77	18	0	0	22	38	16	16	15	23
10	B	0	1	0	0	98	79	0	0	63	65	35	34	21	55
11	A					100	70	100	77	68	61	67	81	71	93
11	B	1	1	0	0	0	0	0	0	0	30	11	11	16	4
12	A					100	82	100	74	83	66	70	84	57	77
12	B	1	1	0	0	100	0	100	12	24	30	46	40	53	37
13	A					67	34	0	0	35	45	21	21	14	8
13	B	0	1	1	0	70	35	0	0	35	46	21	21	21	64
14	A					97	91	0	0	59	70	38	37	21	22
14	B	0	1	1	0	91	77	0	0	51	64	34	33	20	18
15	A					100	1	100	99	10	31	46	68	71	76
15	B	1	1	1	0	100	1	100	100	10	31	46	68	71	76
16	A					100	23	100	100	35	40	53	75	80	92
16	B	1	1	1	0	0	0	0	0	0	30	11	11	14	4
17	A					100	19	100	37	18	39	51	54	66	69
17	B	0	0	0	1	100	0	100	36	17	30	46	48	65	46
18	A					100	11	100	40	17	35	49	52	55	63
18	B	0	0	0	1	100	7	100	37	9	33	48	50	56	66
19	A					100	74	100	7	65	63	68	60	55	53
19	B	1	0	0	1	100	72	100	7	65	61	67	59	55	49
20	A					100	92	100	8	72	70	73	66	57	41
20	B	1	0	0	1	100	88	100	4	53	68	72	63	56	71
21	A					100	47	100	40	45	51	60	63	58	80
21	B	0	0	1	1	100	9	100	21	10	34	48	46	59	18
22	A					100	100	100	89	83	74	76	94	47	60
22	B	0	0	1	1	100	8	100	0	15	34	48	39	53	42
23	A					5	0	0	0	0	31	11	11	15	3
23	B	1	0	1	1	91	79	0	0	72	65	35	34	22	39
24	A					100	66	100	4	63	59	66	57	57	44

Clip	Ball	Outcome				Cause					Model				Rating
		both	only A	only B	neither	difference	whether	how	sufficient	robust	CSM _W	CSM _{WH}	CSM _{WHS}	Heuristic	
24	B					100	94	100	22	79	71	74	71	54	73
25	A	0	1	0	1	100	25	100	21	26	41	53	50	69	43
25	B					100	74	100	54	65	62	68	75	56	73
26	A	0	1	0	1	100	6	100	3	9	33	47	39	60	39
26	B					100	87	100	35	54	68	72	73	46	69
27	A	1	1	0	1	100	97	100	52	97	73	75	81	67	80
27	B					0	0	0	0	0	30	11	11	17	6
28	A	1	1	0	1	100	90	100	22	80	69	73	69	79	89
28	B					0	0	0	0	0	30	11	11	12	5
29	A	0	1	1	1	100	58	100	24	44	56	63	61	66	47
29	B					100	63	100	24	38	58	65	62	54	67
30	A	0	1	1	1	100	57	100	29	49	55	63	62	62	58
30	B					100	46	100	24	39	51	60	58	63	56
31	A	1	1	1	1	100	2	100	4	3	31	46	38	63	44
31	B					100	4	100	4	4	32	47	39	53	46
32	A	1	1	1	1	0	0	0	0	0	30	11	11	16	5
32	B					100	75	100	66	73	63	68	79	65	71

Table A2

Results of individual participant model comparison. Note: r = Pearson's correlation coefficient, RMSE = root mean squared error. The best-fitting model for each participant was chosen using approximate leave-one-out crossvalidation. W = whether-cause, H = how-cause, S = sufficient-cause.

participant	r_w	r_{wh}	r_{whs}	RMSE _w	RMSE _{wh}	RMSE _{whs}	best model
1	0.31	0.38	0.47	29.91	29.00	27.91	whs
2	0.24	0.77	0.77	39.12	27.82	27.43	whs
3	0.39	0.78	0.79	38.75	28.04	27.06	whs
4	0.41	0.54	0.61	43.73	40.80	39.08	whs
5	0.42	0.50	0.51	39.68	37.81	37.40	whs
6	0.21	0.54	0.54	40.70	36.30	35.99	whs
7	0.52	0.62	0.64	32.76	29.87	29.06	whs
8	0.41	0.63	0.67	38.27	33.26	32.02	whs
9	0.42	0.71	0.75	35.76	28.39	26.63	whs
10	0.31	0.52	0.55	29.01	26.24	25.66	whs
11	0.52	0.56	0.60	33.59	32.23	31.20	whs
12	0.55	0.55	0.65	33.21	32.54	30.21	whs
13	0.58	0.70	0.73	29.74	25.59	24.50	whs
14	0.45	0.47	0.50	33.67	32.88	32.27	whs
15	0.24	0.37	0.40	39.54	38.12	37.50	whs
16	0.62	0.65	0.73	40.91	38.61	35.67	whs
17	0.45	0.67	0.71	28.26	23.70	22.27	whs
18	0.32	0.36	0.38	37.76	37.08	36.65	whs
19	0.38	0.60	0.62	33.36	29.29	28.68	whs
20	0.57	0.60	0.66	36.70	35.33	33.55	whs
21	0.43	0.49	0.54	31.05	29.79	28.84	whs
22	0.51	0.66	0.69	33.83	29.68	28.62	whs
23	0.33	0.44	0.51	38.80	37.09	35.88	whs
24	0.46	0.62	0.68	34.13	30.18	28.49	whs
25	0.63	0.66	0.70	30.41	28.88	27.66	whs
26	0.60	0.72	0.76	40.17	35.05	32.98	whs
27	0.46	0.50	0.62	43.42	41.78	39.22	whs
28	0.33	0.45	0.47	29.19	27.59	27.29	whs
29	0.50	0.59	0.64	38.89	36.09	34.73	whs
30	0.22	0.71	0.71	32.92	24.91	24.51	whs
31	0.49	0.59	0.68	37.52	34.81	32.17	whs
32	0.39	0.60	0.69	40.56	36.03	33.32	whs
33	0.51	0.55	0.62	38.76	37.04	35.38	whs
34	0.28	0.51	0.54	44.27	40.70	39.80	whs
35	0.19	0.33	0.32	34.41	33.26	33.23	wh
36	0.23	0.84	0.83	45.20	29.54	29.49	whs
37	0.55	0.62	0.65	36.21	33.72	32.64	whs
38	0.52	0.60	0.64	36.18	33.66	32.38	whs
39	0.49	0.77	0.77	29.40	21.81	21.74	wh
40	0.49	0.78	0.80	32.19	23.68	22.50	whs
41	0.35	0.50	0.57	32.72	30.30	28.86	whs