A counterfactual simulation model of causal judgment

Tobias Gerstenberg*
MIT

Noah D. Goodman
Stanford University

David A. Lagnado
University College London

Joshua B. Tenenbaum
MIT

Abstract

How do people make causal judgments? We propose a counterfactual simulation model (CSM) of causal judgment. The CSM predicts causal judgments by comparing what actually happened with counterfactual simulations of what would have happened in relevant contingencies. It postulates different aspects of causation that capture whether the cause made a difference to *whether* and *how* the outcome occurred, and if the cause was *sufficient* and *robust*. We test the CSM in three experiments that ask participants to make judgments about dynamic collision events. Experiment 1 shows that there is a very close quantitative mapping between causal judgments, and participants' belief that the outcome would have been different without the cause. Experiment 2 establishes that postulating counterfactual contrasts is necessary for explaining causal judgments by showing how participants' judgments differ dramatically between pairs of situations in which what actually happened was identical, but what would have happened differed. Experiment 3 features two candidate causes and shows how 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, deal with causation by omission, and capture judgments of physical stability.

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

*Corresponding author: Tobias Gerstenberg (tger@mit.edu).

We thank all the whether-causes without which this paper wouldn't have happened, and the how-causes who have helped to improve it. In particular, we thank Christopher Hitchcock, Christos Bechlivanidis, Jonas Nagel, Jonathan Kominsky, Jonathan Phillips, Jonathan Schaffer, Joseph Halpern, Joshua Hartshorne, Joshua Knobe, Julian DeFreitas, Kevin Smith, Laurie Paul, Liang Zhou, Max Kleiman-Weiner, Members of CoCoSci, Members of Lagnado Lab, Michael Waldmann, Ned Hall, Nori Jacobi, the participants of the Hoboken workshop, 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), Pascale Willemsen, Peter Battaglia, Phillip Wolff, Pooja Paul, Ralf Mayrhofer, Richard Holten, Shaun Nichols, Simon Stephan, and Tomer Ullman. This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216.

Contents

To do list

- **modeling**:
  - clip 3 turns out to be overdetermined given the setup of the walls → shall we change slightly so it's not?
  - robustness doesn't matter since it's very highly correlated with whether-causation
  - difference-making: is there any gradation?
  - treat model as running on individual participant level:
    - run the model for each participant; assuming 10 simulations for each test per participant (and maybe 3 simulations for difference-making)?!
    - then aggregate the predictions across participants for different versions of the model
    - use regression to adjust the weights for each factor
  - how to apply noise?
    1. current version: apply noise to everything else
    2. version to test: apply noise only to the object that was affected by the collision
    3. possible hybrid: small noise to everything, larger noise to the object that would have participated in the collision
    - check whether this makes any difference
    - double-prevention case suggests that the 2nd version is more appropriate
- make figures with stimuli bigger
- have a table in the general discussion with alternative models and discuss them qualitatively
- **general discussion**:
  - robustness
    - 2 vs. 3 balls in a causal chain (robustness a good way of explaining why causal responsibility is lower)
  - how-causation
    - affecting vs. helping to bring about (in a better way)
  - almost
    - how-cause can be an almost-preventer
- **discussion**:
  - talk about mini experiments in the
  - how-causation: did it bring it about in a better way
  - how-cause vs. almost prevented
  - talk about almost-caused and almost-prevented
  - present these cases: causing vs. almost preventing
  - be clear about the limitations: and talk about how to address in the future

  - how-causation
  - almost
  - robustness
- **materials**:
  - counterfactual clips for all of the videos as generated with python
  - ask kevin about ugly teleport in pygame (due to rotation of png)
  - provide figures with individual data for repository online
- motivate the number of samples based on eye-tracking paper
- don't say much about process vs. computational
- potential people to give feedback:
  - ned hall
  - jonathan livengood
  - chris hitchcock
  - joe halpern
  - jonathan phillips
  - thomas icard
  - jonathan kominsky
  - fiery cushman
  - josh knobe
- paper:
  - typeset the aspects consistently (always smallcaps?)

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 had forgotten 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* things happened. They explain the outcomes of interest 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. It is the glue of the universe that holds events together. Knowing how the world works allows us to make predictions about the future, reason about the past, and choose actions that get us what we want.

Despite the importance of causality, or maybe because of it, attempts to provide a unifying account of how people make causal judgments have proven elusive. 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 many to endorse a pluralistic view, postulating two,

or more fundamentally different concepts of causation (e.g. Cartwright, 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 particular events. The CSM draws from philosophical theories about the nature of causation (Beebee, 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 are about difference-making (Woodward, 2003). Only things that made a difference are causes. Second, to understand causal judgments about particular events ("Joe's shot killed Bill.") rather than general causal relationships ("Shooting kills people."), we need to analyze difference-making in terms of counterfactual contrasts (Lipe, 1991), by comparing what actually happened with what would have happened if the candidate cause had been absent (or different). Third, there are a number of ways in which a candidate cause can make a difference to the outcome. For example, a cause can make a difference to *whether* the outcome occurred, and to *how* it came about (Glymour et al., 2010; Hitchcock, 1996; Lewis, 2000; Schaffer, 2005; **?**). Capturing these different *aspects* of causation is critical to understanding people's causal judgments. The CSM unifies existing theories of causation by showing how the different aspects of causation can each be expressed in terms of counterfactual contrasts operating over the same intuitive model of the situation.

As a general model of causal judgment, the CSM makes predictions within any domain that can be represented as a generative model (Chater & Oaksford, 2013). As a case study, we focus on people's causal judgments about dynamic collisions between billiard balls. This domain is sufficiently rich to model a wide range of situations that have been discussed in the literature, such as overdetermination, joint causation, preemption, and double prevention. The CSM is the first model to make accurate, quantitative predictions about people's causal judgments across a wide range of dynamic physical scenarios.

The rest of the paper is organized as follows. First, we will motivate the problem of causal judgment and discuss some of the major philosophical theories of causation. Then we provide an overview of the empirical landscape that psychological work on causal judgments has painted thus far. Afterwards, we discuss the counterfactual simulation model in detail. We will then show evidence for the model coming from three experiments, starting with a simple setup, and moving on to increasingly complex ones that feature multiple candidate causes. We conclude by discussing remaining challenges and future directions.

The philosophy of causation

Consider the following scenario: Billy threw a stone at a bottle. The stone hit the bottle, and the bottle shattered. The stone's hitting the bottle (C) caused the bottle to shatter (E). 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 of causation, C was a cause of E if C and E were connected via a spatiotemporally continuous process. Processes are defined in terms of a transfer of some quantity such as physical force (Aronson, 1971; Dowe, 2000; Fair, 1979; Machamer, Darden, & Craver, 2000; Salmon, 1984, 1994; Waskan, 2011). For our example, Billy's throw is

identified as the cause of the bottle's shattering since 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 in some way dependent on C. The notion of dependence has been captured 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 would 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 qualifies as the cause of the bottle's shattering because the bottle would not have shattered if we had intervened in the actual course of events and made it such that Billy did not throw the stone.

A key advantage of interventionist theories is that the criteria for whether or not an event qualifies as an actual cause of a particular outcome are defined precisely. These theories represent the situation in terms of causal networks that capture the causal dependencies between variables representing the events of interest. A set of structural equations then specifies the exact way in which the different variables depend on each other. Given a choice of model (which includes the variables used to represent the scene, their possible values, and the structural equations relating the variables), these *structural theories* of causation say for each variable in the causal network, whether it was a cause of the outcome of interest (Halpern, 2016).

For simple scenarios like the one above, both process theories and dependence theories generally yield the same verdict. However, in other situations, the verdicts of the two frameworks come apart. Consider a modification of the scenario in which Billy (C_1) and Suzy (C_2) both throw stones at the bottle (cf. 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. Thus, the shattering of the bottle in that situation was causally overdetermined. Was C_1 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 situations of overdetermination. In this modified scenario, there was a spatiotemporally continuous process from each of Billy's and Suzy's throws. Each stone transfered force to the bottle. Hence, both throws are deemed causes of the bottle's shattering.

Dependence theories, in contrast, have trouble with situations of overdetermination. The bottle would still have shattered if either Billy or Suzy hadn't thrown their rock. Thus, according to a simple counterfactual criterion, neither C_1 nor C_2 are deemed causes of the bottle's shattering.

Several attempts have been made to handle the problem of overdetermination (Lewis, 1973). One solution is to say that while it is true that neither C_1 nor C_2 individually were causes of E, they both together caused E to happen (Halpern, 2016). Another strategy is to increase the granularity of the outcome event. Accordingly, while it is true that neither C_1 nor C_2 were causes of the bottle's shattering E (broadly construed) each was a cause of the exact way in which the bottle shattered ΔE (finely construed). If either C_1 or C_2 hadn't occurred than the bottle would have shattered differently from how it actually did (Lewis, 2000). Whether C_1 qualifies as a cause of E thus depends on what counterfactual

contrast for E we are considering: the bottle not shattering ($\neg$E), or the bottle shattering differently (ΔE$'$) (cf. Schaffer, 2005).

Finally, one can also relax the simple counterfactual criterion by not only considering whether E was counterfactually dependent on C in this particular situation, but also whether E would have been counterfactually dependent on C in other situations that could have happened (Halpern & Pearl, 2005; Hitchcock, 2001; Woodward, 2003; Yablo, 2002). For example, in the overdetermination scenario, Billy's throw was not pivotal for the bottle's shattering in the actual situation because of Suzy's throw. Billy's throw was not pivotal because the bottle would have shattered even if Billy had not thrown his stone. However, in the counterfactual situation in which Suzy hadn't thrown, Billy's throw would have been pivotal for the bottle's shattering.

While process theories handle situations of overdetermination with ease, they have trouble with other kinds of situations that we will discuss below. Furthermore, defining in non-causal terms what distinguishes a causal process from other kinds of processes has proven difficult (Ehring, 1986; Hitchcock, 1995). If one billiard ball C hits another billiard ball E and E subsequently moves, we say that C caused E to move. But what is the relevant causal process that connects C and E? Intuitively, it is the transfer of momentum from C to E that did the causing, and not the chalk that was transferred as well. It looks like counterfactual criteria are necessary to determine which process was causal: whereas removing the chalk from ball C would have still resulted in E moving, "removing" the momentum from ball C would have led to E staying put (cf. Woodward, 2011a). Process theories of causation generally aim to reduce causality to what actually happened and do away with "esoteric speak" about possible counterfactual worlds (e.g. Salmon, 1994). While having a theory of causation that is grounded merely in what actually happened may be a desirable metaphysical goal (Paul & Hall, 2013), we will see below that counterfactuals are necessary to explain people's causal judgments.

The psychology of causal judgment

In this section, we will first give a brief overview of the empirical landscape, focusing on work that investigated how people's causal judgments are affected by information about causal processes as well as information about counterfactual dependence. While most of the work on causal judgment to date has been largely qualitative, there have also been some attempts to develop formal theories of causal judgment. We will discuss the *force dynamics model* (Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010) which is rooted in causal process theories, and a framework for analyzing causation based on structural causal models (Halpern, 2016; Halpern & Pearl, 2005) which is inspired by dependence theories of causation. Finally, we will foreshadow how the counterfactual simulation model proposed here combines key insights from both process and dependence theories to arrive at a more unified model of causal judgment.

Empirical work on causal judgment. Philosophical debates on how to best analyze causation have heavily influenced psychological work on how people make causal judgments (Hitchcock, 2012; Woodward, 2011b). Psychologists have identified to what extent different factors such as covariation information, counterfactual dependence, or information about causal processes influence people's causal judgments (cf. Einhorn & Hogarth, 1986).

Evidence about how much information about processes versus counterfactual dependence affects participants' causal judgments has been mixed.

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 in line with the predictions of process rather than regularity theories of causation. Regularity theories predict that people learn about causal relationships through considering covariation and spatiotemporal contiguity (Cheng, 1997; Cheng & Novick, 1992; Hume, 1748/1975). Shultz (1982) found that participants' judgments were more strongly affected by the presence of a plausible mechanism as opposed to dependence information such as the timing of events.

Mandel (2003) conducted a number of scenario-based experiments to investigate the relationship between causal judgments and counterfactual thinking. For example, one experiment featured a scenario in which Mr. Wallace, who was highly influential in the organized crime scene, is first lethally poisoned over lunch by Mr. Vincent. On his way to another business meeting, Mr. Wallace's van is pushed off the side of the road by Mr. Bruce. The car exploded and the coroner's report states that Mr. Wallace had received fatal burns in the car explosion. Mandel (2003) found that participants' causal and counterfactual selections came apart: when asked what would have needed to be different in order to undo Mr. Wallace's premature death, participants tended to select Mr. Wallace's involvement in crime. When asked to say what caused Mr. Wallace's death, participants tended to select the car crash. Mandel explains this dissociation by postulating that counterfactual judgments focus on events that were necessary for the outcome to occur, whereas causal judgments focus on events that were sufficient under the circumstances – events that pick out the actual process that caused the outcome (cf. Mandel & Lehman, 1998).

Mandel's (2003) results show that there are situations in which explicit causal and counterfactual judgments dissociate. These results don't show, however, that causal and counterfactual judgments are unrelated. Indeed, as we will suggest below, counterfactuals play a crucial role in defining what it means for something to have been sufficient in the circumstances, and for analyzing *how* the outcome came about.

Additional evidence for the role of causal processes in people's cause and prevention judgments comes from a series of vignette studies by Walsh and Sloman (2011). In one scenario, Frank accidentally kicks a ball toward a neighbor's house. Sam, his friend, is initially blocking the ball's path but gets distracted and steps out of the way. As a result, the ball hits the neighbor's window and smashes it. When asked whether Frank (who kicked the ball) caused the window to smash, 87% of participants answered positively. In contrast, only 24% of participants agreed that Sam (who stepped out of the way) caused the window to smash.

Lombrozo (2010) used similar scenarios and looked at how manipulating whether or not the agents acted intenionally affects people's causal judgments. The "transference cause" (i.e. Frank's kicking the ball) that directly influenced the outcome was generally judged as causal no matter whether it had brought about the outcome intentionally or by accident. In contrast, the"dependence cause" (i.e. Sam's going out of the way in the previous scenario) was rated lower when it was accidental compared to intentional. Lombrozo argues that intentions matter for causal judgments because they affect the perceived robustness between cause and effect (cf. Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015;

Woodward, 2006). A causal relationship is robust to the extent that it would have held even if the circumstances had been somewhat different (cf. Heider, 1958). For example, if Sam wanted for the window to break, he would have made sure to get out of the way even if Frank had shot the ball differently. However, if Sam's going out of the way was accidental, then a different shot by Frank might have hit Sam instead of the window.

So far, we have seen evidence that people care about process information when making causal judgments: this can lead to dissociations between causal and counterfactual judgments, and a preference for transference over mere dependence causes of an outcome. However, there is also evidence that counterfactual dependence sometimes matters more than causal processes.

Chang (2009) directly pitted process theories and counterfactual theories against each other. In his experiments, a toy train ran into a card house causing the cards to fall down. In some situations, an agent pushed the train and thus the action was physically connected to the outcome. In other situations, the train was already moving and the agent opened a gate that would otherwise have blocked the train. While for both situations, the outcome was counterfactually dependent on the agent's action, only the former situation involved a direct transmission of force between action and outcome. In order to manipulate counterfactual dependence, the outcome in some situations was overdetermined by introducing another train that was approaching 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 mostly determined by whether or not the outcome was counterfactually dependent on the agent's action. 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 made no difference.

Theories of causal judgment. Research into causal judgments has suffered from a lack of formally specified theories. The studies we have discussed so far have relied on comparing qualitatively, whether causal judgments depending on the way in which the cause brings about the effect. Rather than defining what makes for a causal process, these studies have relied on our intuitive understanding of what it means for a cause to directly affect the outcome. We will now discuss two models of causal judgment: one rooted in process theories, and one rooted in dependence theories.

TG: might want to weaken this

Force dynamics model. 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). In contrast to dependence theories which have almost exclusively focused on causation and prevention, the FDM captures a number of causal terms including "caused", "helped", "prevented", and "despite". Each term is analyzed in terms of a different configuration of forces that characterizes the interaction between agent (A) and patient (P) with respect to some endstate (E).

For example, the FDM predicts that an agent *caused* a patient to reach a certain endstate if, a) the patient did not have a tendency to reach the endstate (i.e. P's force vector did not point toward E), b) 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 c) the patient did in fact reach the endstate. To make this more concrete, consider a situation in which a boat (the patient) cruises on water 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 the fans *caused* the boat to hit the cone in this situation.

If, in contrast, the boat was already headed toward the cone and the fans blew from behind toward the cone, the FDM predicts people will say that the fans *enabled* (or *helped*) the boat to reach the cone. Wolff (2007) reports several experiments that show how the FDM predicts participants' selection of causal terms very accurately.

The FDM analyzes causation without counterfactuals. Different causal terms are explained directly via the force configurations they map onto. Wolff discusses that the force representation supports counterfactual simulation. For example, one could imagine what would happen if the patient's force had been absent or different. However, counterfactual contrasts are not required to explain the different causal terms. The force representation is primary, and both causal as well as counterfactual judgments derive from it. While the model considers the patient's tendency at the time of the causal interaction, this tendency is explicitly not defined in counterfactual terms. So, rather than thinking of the tendency in terms of what would have happened to the patient if the agent's force had been absent, the patient's tendency is defined as the direction of force at the time of interaction.

In recent years, (Wolff et al., 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 difficult to accommodate by process theories of causation (McGrath, 2005; Schaffer, 2000a). In order to deal with omissions, the FDM moves beyond standard process theories of causation by allowing for causation to take place even when there was no direct transmission between cause and effect (cf. Dowe, 2000, 2001).

Wolff et al. (2010) claim that causation by omission is always embedded within a causal structure of double prevention. In a double prevention scenario, A prevents B which would otherwise have prevented C from happening. The example that Wolff et al. give is that of a man (A) who pushes the jack aside that holds a car off the ground (B) resulting in the car falling to the ground (C). Here, the man prevented the jack from preventing the car from falling to the ground. The man's removal of the force exerted by the jack on the car, caused the car to fall down. Note, however, that there is no direct transmission of force from A to C. The man only exerts force to remove the jack which leads to the car falling down. He doesn't exert force on the car directly.

While Wolff et al. regard counterfactuals as necessary for handling causation by omission, they maintain that for assessing simple causal relations counterfactuals are not required. In contrast, we believe that causal judgments are intimately linked to counterfactuals, and that even understanding simple causal judgments requires considering counterfactual contrasts.

Structural causal model. To discuss how actual causation can be captured formally within the framework of dependence theories, we will focus on an account developed by Halpern (2016) which we will refer to as a *structural causal model* (SCM) of causal judgment (for related accounts, see Hitchcock, 2001; Woodward, 2003; Yablo, 2002).

According to the SCM, events of interest are represented as variables, and the causal relationships between events are defined by structural equations relating the variables. Let

us illustrate the account via an example. Suzy and Billy throw stones at a bottle. If either of them hits the bottle, the bottle shatters. Equation 1 shows a simple structural model representing the situation just described. We model Billy's and Suzy's hitting the bottle as BH and SH, respectively. The equations further say that the bottle shatters (BS) if either Billy or Suzy (or both) hit the bottle as indicated by the logical disjunction ($\vee$).

$$\begin{aligned} &\mathrm{BH} \\ &\mathrm{SH} \\ &\mathrm{BS} = \mathrm{BH} \ \vee \ \mathrm{SH} \end{aligned} \tag{1}$$

For simplicity, let us assume that each variable is binary. So, Billy can either hit the bottle (BH = 1) or miss it (BH = 0). The relations between the variables express how the world works. For example, the model expresses that if Billy's stone hits the bottle (BH = 1), then the bottle shatters (BS = 1) no matter what Suzy does. However, the structural equations do not yet answer the question of whether in a particular situation, one variable caused another. Was Suzy's hitting the bottle (SH = 1) a cause of the bottle's shattering (BS = 1) in a situation in which Billy also hit the bottle (BH = 1)?

A lot of work has gone into finding the right criteria so that the model's answer of whether one variable caused another agrees with our intuition (Halpern & Pearl, 2005; Hitchcock, 2001; Woodward, 2003; Yablo, 2002). What all accounts have in common is that they take the simple test for counterfactual dependence as a starting point. Accordingly, 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 in this framework counterfactual dependence is sufficient for causation, it is not necessary. In our example, the bottle would still have shattered even if Suzy hadn't hit it because Billy hit the bottle as well. However, Suzy's hitting the bottle was clearly a cause of the bottle shattering.

To accommodate this intuition, the SCM employs a more sophisticated definition of actual causation. According to this definition, a variable can qualify as a cause of another variable even if they were not counterfactually dependent in the actual situation, as long as a situation is possible in which they would have been dependent. For example, the bottle shattering would have been counterfactually dependent on Suzy's throw if Billy had missed the bottle (BH = 0). Getting the definition of actual causation just right such that it agrees with people's intuitions across a range of different situations has proven a difficult task.

The SCM models counterfactuals as interventions on the causal system. Pearl (2000) introduced the *do*() operator to model inferences based on interventions. Like an idealized experiment, the *do*() operator modifies the structural equations that express how the system works by removing the equation of the variable that was intervened on, and simply setting that variable to the desired value. As a consequence, inferences based on observations are different than inferences based on interventions.

For example, imagine that you *observe* a shattered bottle but you don't know whether Billy or Suzy threw their rock. Given the causal structure of the situation, you can infer from the fact that BS = 1 that either Billy or Suzy (or both) must have thrown their rock (i.e. the posterior $P(\mathrm{BH}, \mathrm{SH}|\mathrm{BS} = 1)$ is greater than the prior $P(\mathrm{BH}, \mathrm{SH})$). If, in contrast, you imagine an external *intervention* that shattered the bottle (i.e. we replace $\mathrm{BS} = \mathrm{BH} \ \vee \ \mathrm{SH}$ with BS = 1 in Equation 1), then you cannot infer anymore from the fact

that the bottle is now shattered that Billy or Suzy threw their stone (i.e. $P(\text{BH}, \text{SH}|do(\text{BS} = 1)) = P(\text{BH}, \text{SH})$).

Research in psychology has shown that people make different inferences based on observations versus interventions (Meder, Gerstenberg, Hagmayer, & Waldmann, 2010; Sloman & Hagmayer, 2006; Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005), although the evidence is mixed as to whether people reason about counterfactuals in the way that the SCM prescribes (Dehghani, Iliev, & Kaufmann, 2012; Gerstenberg, Bechlivanidis, & Lagnado, 2013; Hiddleston, 2005; Lucas & Kemp, 2015; Rips & Edwards, 2013).

An important limitation of the SCM is that it does not yield any graded judgments of causation. For a given structural model and a setting of the variables, the definition merely says whether a particular variable caused another. It doesn't say how good a cause it was. To remedy this situation, structural equation models have been supplemented with considerations of normality (Halpern & Hitchcock, 2015) to capture the fact that we tend to favor abnormal over normal events as causes (Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard, Kominsky, & Knobe, 2017; Kominsky et al., 2015; Samland & Waldmann, 2015). It has also been proposed that the causal responsibility of a variable increases the closer a situation was in which the variable would have made a difference to the outcome (Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2010; Lagnado, Gerstenberg, & Zultan, 2013; Zultan, Gerstenberg, & Lagnado, 2012).

Bridging process and dependence accounts of causation. Both process and dependence accounts of causation capture key aspects of how people make causal judgments (cf. Ney, 2009; Woodward, 2006). The CSM aims to combine the best of both worlds. 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 to the outcome in one way or another are considered causal candidates of the outcome. And since we are dealing with 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).

However, dependence theories have traditionally focused on what we will refer to as the "whether-aspect" of causation (Ahn & Kalish, 2000; Mandel, 2003). For example, the variables in structural models tend to be specified on a coarse level of granularity that merely denote the presence versus absence of the event of interest (Woodward, 2015). Process theories, in contrast, have focused on what we will call the "how-aspect" of causation. (Wolff et al., 2010, p. 215) argue that "people simulate the processes that produce causal relationships rather than simply specifying the dependencies that hold between one event or state and another". We too believe that people represent the situation in terms of a generative model which describes the processes that lead to the production of the outcome.

The process of mental simulation plays a central role in our account (Hegarty, 2004; Kahneman & Tversky, 1982). Kahneman and Tversky (1982, p. 201) argue that "there appear to be many situations in which questions about events are answered by an operation that resembles the running of a simulation model. The simulation can be constrained and controlled in several ways: The starting conditions for a 'run' can be left at their realistic default values or modified to assume some special contingency". The counterfactual simulation model we propose below provides a concrete implementation of this idea. We assume that people use their mental model of the situation not only to predict what actu-

ally happened, but also to simulate what would have happened in different counterfactual contingencies (Chater & Oaksford, 2013; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Kahneman & Tversky, 1982; Roese, 1997). A detailed generative model of the situation allows us to capture both whether candidate causes made a difference to *whether* the outcome occurred as well as to *how* it came about (Lewis, 2000; Woodward, 2011a).

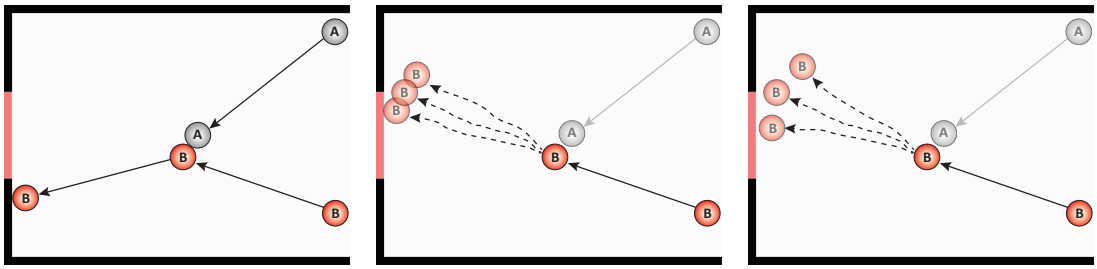
Unlike the force dynamics model (Wolff, 2007; Wolff et al., 2010), the counterfactual simulation model grants forces no special status. While our understanding of forces is central to making causal judgments in the physical world, we do not believe that the concept of force is fundamental to making causal judgments in other domains such as when we reason about other agents (cf. Baker, Saxe, & Tenenbaum, 2009; Heider, 1944). When we try to figure out why Joe suddenly turned around and walked back home, we think about his beliefs, desires, and intentions as candidate causes. We simulate Joe's decision process by making use of our intuitive theory of mind (Gerstenberg & Tenenbaum, 2017). Similarly, when we try to make sense of whether the fall of Lehman brothers caused the financial crisis, we use our intuitive theory of the financial system to simulate what might have happened if Lehman brothers had been saved. Forces don't play an important part for how we think about the financial system. However, we may have a more abstract economic model that describes how the financial system works, and we can simulate the consequences of different counterfactuals on this system.

In sum, the CSM unifies process and dependence theories by assuming that people represent their knowledge about how the world works as generative models that capture the causal processes of how outcomes are produced. Instead of postulating fundamentally different *concepts* of causation that are associated with processes or dependence (cf. Hall, 2004), the CSM posits different *aspects* of causation which are defined as counterfactual constasts on the generative model. These aspects express the different ways in which a cause can make a difference to the outcome, such as *whether* or *how* it occurred.

The Counterfactual Simulation Model

The *counterfactual simulation model* (CSM) predicts causal judgments about objects (or agents) for particular events, such as "the stone caused the window to shatter" or "Tim caused the vase to break".[1] The CSM is a general model of causal judgment and applies to different domains of interest. Here, we illustrate the workings of the model by focusing on people's causal judgments about video clips that show dynamic interactions in a physical domain. Specifically, participants' task in our experiments was to judge whether one billiard ball caused another ball to go through a gate, or prevented that ball from going through the gate. Figure 1a shows a diagrammatic illustration of one clip. In Experiment 3, participants saw more complex interactions between three billiard balls that allowed us to reconstruct many of the situations for which process theories and dependence theories of causation yield

[1] In line with most philosophical accounts, we believe that causal relata are best charaterized as *events*. However, it is often more natural to talk about objects or agents having caused an outcome, rather than the events that they participated in. For example, it is more natural to say that "the stone caused the window to shatter" rather than "the stone's hitting the window caused the window to shatter". As we will see below, it is often also more natural to express counterfactual operations on candidate objects rather than on the events that they participated in.



(a) actual situation

(b) three counterfactual simulations with small degree of noise

(c) three counterfactual simulations with large degree of noise

Figure 1. a) Illustration of what actually happened. b) and c) Illustrations of approximate simulations of where 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.

different predictions, such as situations of joint causation, overdetermination, preemption, and double prevention.

DL: This makes our studies sound too narrow - can we add sentence clarifying that such contexts are very generic examples of physical causation (perhaps refer to Hume - who I think considered such examples as the ultimate causal relations)

TG: not sure whether i've addressed this sufficiently

TG: maybe add something along the lines of: Our model is a general model of causation. However, we focus on the domain of physical judgments since this is the domain in which the support for process model is strongest. We show that counterfactual contrasts are critical even in this domain ...

The CSM explains people's causal judgments by assuming that people use their intuitive understanding of the domain to compare what actually happened with what would have happened in different counterfactual situations.[2] The model operates in two steps: first, it uses a fine-grained test of difference-making to identify all candidate causes. In the second step, it then predicts for each candidate cause, to what extent it caused the outcome. In other words, the first step filters out which candidates were "a cause" of the outcome and which ones were not. The second step, determines for each candidate cause to what extent it was "the cause" of the outcome (Hart & Honoré, 1959/1985; Hesslow, 1988; Hilton, 1990).

[2]We assume that people's intuitive domain knowledge can be represented as a generative model. Because we assume that people have a causal theory that describes how the world works we do not provide a model of causal judgment that tries to reduce causality to something else (for a philosophical attempt to provide a reductive analysis of causation, see Lewis, 1973, 2000). Further, we assume that our participants already have a fairly sophisticated understanding of how the physical world works. We do not model the process of how people arrive at this understanding (Gerstenberg & Tenenbaum, 2017; Lake, Ullman, Tenenbaum, & Gershman, 2016; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Wellman & Gelman, 1992).

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. We have seen above that process theories and dependence theories employ different criteria for fixing causation. 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. Furthermore, we define the outcome event on a fine level of granularity that specifies not only whether or not the outcome happened, but also captures the 'when' and 'where' of what happened (cf. 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}) for how event Δe came about as

$$P_{DM}(C \rightarrow \Delta e) = P(\Delta e' \neq \Delta e | S, remove(C)). \tag{2}$$

In words, to determine whether C was a difference-maker $P_{DM}(C \rightarrow \Delta e)$, we first take 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 the motion paths of the different balls. We then consider the counterfactual situation in which the candidate cause had been removed from the scene $remove(C)$, and evaluate 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 arrow $\rightarrow$ expresses the direction of the causal relation. The Δ indicates that we construe the outcome event finely. That is, we not only care about whether the outcome happened or didn't happen, but how it happened on a finer level of granularity that includes information about exactly where and when the outcome happened.[3] The *remove*() operation is analogous to Pearl's (2000) *do*() operator discussed above. However, instead of implementing interventions by setting a variable in a system of structural equations, we intervene in the physics engine that generated the observed clip by removing the candidate causal object.

To determine whether C was a difference-maker, the model runs a small number N of counterfactual simulations in which C was removed from the scene. C qualifies as a difference-maker if $\Delta e' \neq \Delta e$ in at least one of the simulations. In some situations, determining whether a candidate cause was a difference-maker is trivial. For example, whenever the cause directly collided with the target, it qualifies as a difference-maker (see Figure 3a). However, in other situations, it is more difficult to assess whether the cause was

[3] We 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 our intuition is still that it 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.

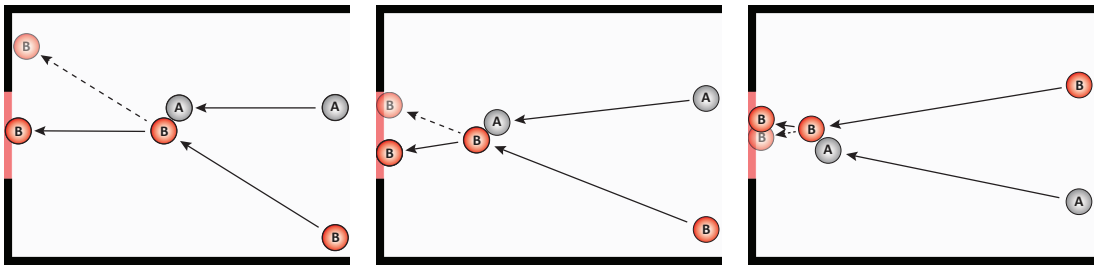
a difference-maker. For example, the situation shown in Figure 3b shows a case of double prevention: ball B prevents ball A from preventing ball E from going through the gate. Here, ball B was a difference-maker of E's going through the gate even though it didn't collide with ball E. If ball B had been removed from the scene, than ball A would have knocked ball E out of the way. If C qualifies as a difference-maker, the CSM proceeds to judging the extent to which it caused the outcome.

Modeling counterfactual simulations. 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. There is growing evidence that certain aspects of our intuitive understanding of physics are well-explained by assuming that we have an approximate physics engine in our mind which we can use to simulate what will happen in the future (Battaglia, Hamrick, & Tenenbaum, 2013; Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016; Kubricht, Holyoak, & Lu, 2017; Smith & Vul, 2013), reason about what must have happened in the past (Smith & Vul, 2014), or make inferences about latent physical properties of objects such as mass or friction (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn, Mansinghka, & Griffiths, 2013; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015).

To evaluate difference-making in the CSM, we need to be able to simulate what would have happened if the candidate cause had been removed from the scene. For the experiments reported in this paper, we created the dynamic stimuli using the physics engines Box2D (`http://box2d.org/`) and Chipmunk (`https://chipmunk-physics.net/`). Physics engines are used in video games to generate realistic looking physical interactions. Physics engines have also been proposed as a working hypothesis for how the mind represents the physical world (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). To assess what would have happened in counterfactual situation in which the candidate cause would have been removed from the scene, we can simply delete the ball from the physics simulation and then simulate what would have happened. However, observers don't have direct access to this ground truth, since they only see what actually happened, not what would have happened. They have to rely on their mental simulation of what would have happened. Because people's mental simulations are only approximate, they have some uncertainty about what would have happened.

Different sources of uncertainty enter people's mental simulation of physical events such as perceptual uncertainty about the position of the balls, as well as dynamic uncertainty about how exactly the objects are going to move (cf. Smith & Vul, 2013). For example, in the situation depicted in in Figure 1a it is unclear whether ball B would have gone through the gate if ball A hadn't been present in the scene. We model people's uncertainty in the counterfactual simulation by introducing noise to B's motion path from the point on at which the two balls would have collided. At each step in the physics simulation, we introduce a random perturbation to the direction of B's velocity vector. A free parameter in the simulation model controls the standard deviation of the Gaussian distribution from which we draw 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 counterfactual simulations where a larger degree of noise was added representing a greater uncertainty about what would have happened.

TG: shall we name this parameter?



(a) B would have clearly missed. (b) B would have just missed. (c) B would have gone in.

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.

In the experiments reported below, we always ask one group of participants to make counterfactual judgments about what would have happened if the candidate cause had been removed from the scene, and we use the approximate simulation model described here to capture people's judgments by fitting the noise parameter to the data.

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 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 of causation WHETHER-CAUSATION, HOW-CAUSATION, SUFFICIENT-CAUSATION, and ROBUST-CAUSATION, and discuss them now in turn.

WHETHER-CAUSATION. Consider the three diagrammatic displays of video clips 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 followed if ball A hadn't been present in the scene. In all three situations, the two balls collided and B ended up going through the gate. Since there was a collision between the balls, A trivially qualifies as a difference-maker P_{DM} of 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 the extent to which A's presence made a difference to *whether or not* 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

$$P_W(C \to e) = P(e' \neq e | S, remove(C)). \tag{3}$$

TG: consistenty type-setting for whether-causation and whether-cause

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, unlike when testing for difference-making, the outcome event is broadly construed this time. It only matters whether the outcome event happened or didn't happen – in our case, whether

or not ball B ended up going 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 from what it was in the actual situation. As proposed above, we assume that in order to determine what would have happened if the candidate cause, ball A, had been removed, people make use of their intuitive understanding of the domain and mentally simulate what would have happened in that counterfactual situation (Gerstenberg, Peterson, et al., 2017).

Let us now see to what extent ball A qualifies as a whether-cause of ball B's going through the gate in the different clips shown in Figure 2. In Figure 2a an observer's subjective degree of belief that A was a whether-cause of B's going through the gate $P_W(A \rightarrow e)$ 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, the situation is less clear. Again, ball B actually went through the gate. However, it is less clear what would have happened if ball A had been removed from the scene. Ball B might have gone through the gate even if ball A hadn't been there. Thus, the CSM predicts that $P_W(A \rightarrow e)$ is intermediate in this case. Finally, in Figure 2c, ball B actually goes through the gate and it is clear that B would have gone through the gate even if ball A had been removed from the scene. Thus, $P_W(A \rightarrow e)$ is low in this case.

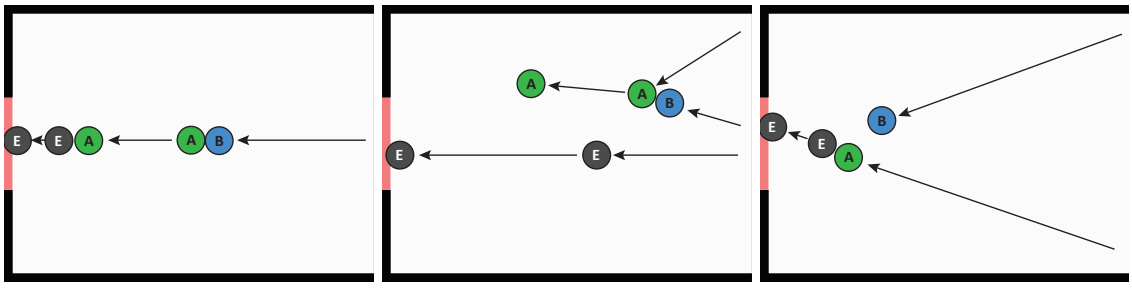
To determine whether-causation we require the counterfactual probability of B's going through the gate in the absence of A. In the experiments reported below, we use two complementary strategies to get these probabilities. First, we directly ask one group of participants to judge whether they think the 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 a number of 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 ended up going through the gate to predict participants' judgment of whether B would have gone through the gate if ball A hadn't been present in the scene. This process of sampling from a generative model by evaluating its simulations nicely dovetails with how (Kahneman & Tversky, 1982, p. 201) conceived of their simulation heuristic: "A simulation does not necessarily produce a single story which starts at the beginning and ends with a definite outcome. Rather, we construe the output of simulation as an assessment of the ease with which the model could produce different outcomes, given its initial conditions and operating parameters. Thus, we suggest that mental simulation yields a measure of the propensity of one's model of the situation to generate various outcomes, much as the propensities of a statistical model can be assessed by Monte Carlo techniques."

TG: be consistent about saying "ball A" or just "A"

HOW-CAUSATION.

TG: make a more explicit link here to Lewis' claim that causation requires a chain of causal dependence (Lewis, 1986), see also (McDermott, 1995); chain of causal dependence is tricky for cases of late preemption

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



(a) Causal chain (b) Double prevention (c) Preemption

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

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 the dichotomy that is often drawn between counterfactual and process 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. Indeed, there is a counterfactual test for determining whether the 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 E lies in front of the gate and ball A lies in the middle of the scene. 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?

TG: consider changing the labels here

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 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 B. Thus, based on whether-causation, we would predict that A has no causal responsibility for E's going through the gate. However, there is clearly a sense in which A contributed to E's going through the gate. Even though A's presence did not make a difference as to whether E went through the gate, it clearly made a difference to how it did so.

How can we capture the intuition that A made a difference to E's going through the gate even though E would have still gone through the gate if A hadn't been present in the scene? 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. We borrow this idea of looking at counterfactual dependence on a finer level of granularity from the philosopher David Lewis (2000) who responded to criticisms of his earlier counterfactual theory of

causation in this way (Lewis, 1973, 1979). However, looking at the outcome event on a finer level of granularity is not enough. Rather than considering what would have happened if the candidate cause hadn't been present in the scene, we need a different counterfactual test. Instead, the CSM considers whether a small perturbation to the candidate cause would have made a difference to the outcome event (finely construed). For example, when considering whether ball A was a how-cause of E's going through the gate in Figure 3a the model simulates a counterfactual situation in which A's position was a little different from what it actually was 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 E's going through the gate.

TG: clear why or do we need to elaborate?

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

$$P_H(C \rightarrow \Delta e) = P(\Delta e' \neq \Delta e | S, change(C)). \tag{4}$$

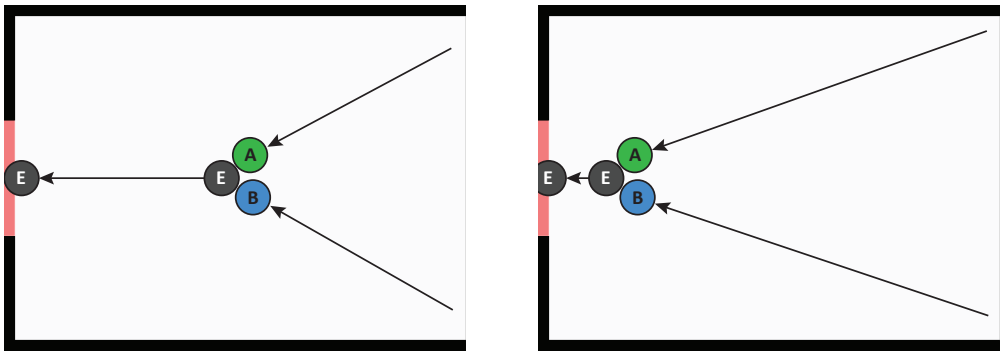
Taking into account what actually happened S, the CSM considers a situation in which the candidate cause was changed $change(C)$ and then simulates whether the event of interest in this situation would have been different from what it actually was $\Delta e' \neq \Delta e$.

The aspect of how-causation captures some of the key intuitions that motivate process theories of causation. It reveals a direct relationship of influence between cause and effect. Indeed, for the kinds of collision events we consider here, how-causation is a simple test for 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. The test for how-causation shows how the intuitive appeal that process theories of causation have by capturing this more direct notion of causal influence, can be expressed as a counterfactual operation (cf. Woodward, 2011a).

How-causation does not generally need to be instantiated via a transfer of force: if Sarah helps John studying for his exam, then she is a how-cause of John's result. She may not have made a difference to whether John passed, but she may have still made a difference to how well he did (cf. McDermott, 1995). Note that even though the tests for difference making (Equation 2) and how-causation (Equation 4) appear similar, they are not redundant. As we will see below a cause can be a difference-maker but fail to be a how-cause.

So far, we have discussed two different aspects of causation: whether-causation and how-causation. For both aspects, the CSM simulates the consequences of a counterfactual intervention on the candidate cause and evaluates whether the outcome would have been different from what it actually was. By considering counterfactuals that affect potential 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 one of the fundamental aspects of causation (e.g. Downing, Sternberg, & Ross, 1985; Hewstone & Jaspars, 1987; Jaspars, Hewstone, & Fincham, 1983; Mackie, 1974; Mandel, 2003; Pearl, 1999; Woodward, 2006). Necessity and sufficiency are often expressed 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



(a) joint causation (b) overdetermination

Figure 4. Two clips in which balls A and B hit the stationary ball E at the same time. In (a) E would not have gone through the gate if either of the balls had been absent from the scene. In (b) E would have gone through the gate even if only one of the two balls had been present.

the effect always occurs in its presence. Here, we are interested in people's causal judgments about particular events: To what extent are balls A and B causally responsible for E's going through the gate in this particular situation? Because we only have a single observation, we cannot use notions of necessity and sufficiency as defined over repeated cause-effect contingencies.

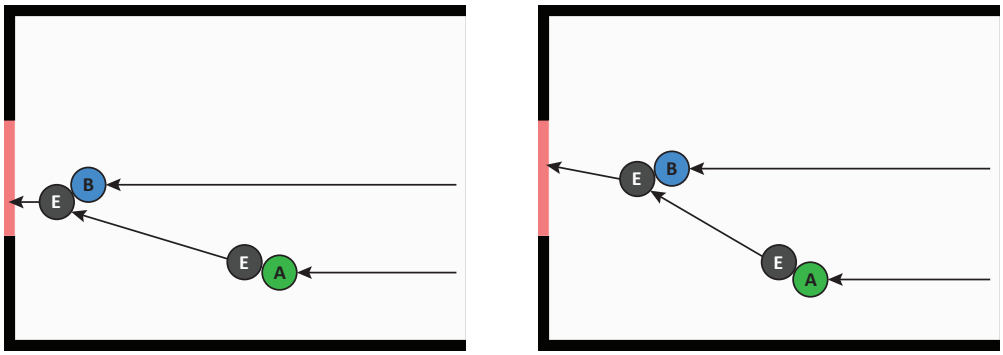
The aspect of whether-causation captures necessity. A candidate cause was necessary if the effect would not have happened, had the cause been removed from the scene (cf. Lipe, 1991). Defining a notion of sufficiency for particular causal relationships is more involved. Previous proposals have in one way or another, relied on more general contingency information when defining sufficiency for particular events (cf. Cheng & Novick, 2005; Icard et al., 2017; Pearl, 1999; Stephan & Waldmann, 2016; Woodward, 2006). Here, we propose a model of sufficiency for particular causal claims such as: "Ball A was sufficient for E's going through the gate in this situation". Our proposal is inspired by the structural-modeling accounts discussed above (Halpern, 2016; Halpern & Pearl, 2005) which tests for counterfactual dependence not only in the actual situation, but also in counterfactual contingencies. The basic idea is the following: when considering whether a candidate cause was sufficient for bringing about the outcome event, the CSM simulates a counterfactual situation in which all other candidate causes were removed from the scene, and then checks whether the candidate cause would have made a difference to the outcome (broadly construed) in that situation.

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

$$P_S(C \rightarrow e) = P_W(C \rightarrow e | remove(\backslash C)). \tag{5}$$

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 $remove(\backslash C)$.

To illustrate the aspect of sufficient-cause, consider the two examples shown in Figure 4. In both examples, ball E is initially at rest and balls A and B hit ball E symmetrically such that E ends up going through the middle of the gate. While both clips are similar in



(a) A was clearly sufficient (b) It is unclear whether A was sufficient

Figure 5. 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. In both situations, A was clearly a whether-cause of E's going through the gate.

terms of what actually happened, their counterfactual profiles are very different. In Figure 4a, both balls were necessary for E's going through the gate. If either of the balls had been removed from the scene, then E would not have gone through the gate. However, neither of the balls was individually sufficient. To check whether ball A was sufficient, we consider a situation in which ball B hadn't been present, and check whether A would have been a whether-cause in this situation. Since E would not have gone through the gate in the situation in which only A was present but not B, A was not sufficient for E's going through the gate. In contrast, in Figure 4b, neither of the balls were individually necessary for E's going through the gate. E would have still gone through the gate even if either A or B had been removed from the scene. However, in this case, both A and B were individually sufficient for E's going through the gate. Ball A is a whether-cause in the counterfactual situation in which B would have been removed from the situation (and vice versa).

In the same way in which we may be uncertain about whether a candidate cause was necessary for the outcome to occur, we may also be uncertain about whether a cause was sufficient. Figure 5 shows two cases which differ in how clear it was whether A was sufficient for E's going through the gate. In both situations, ball E is initially at rest. Ball A first collides with ball E and then ball B collides with E before E goes into the gate. The key difference between the clips is in what would have happened if ball B hadn't been present. In Figure 5a it is relatively clear that ball E would have gone through the gate even if ball B hadn't been there, and A was thus sufficient for E's going through the gate. In contrast, in Figure 5b it is less clear whether ball E would have gone through the gate even if ball B hadn't been present. Accordingly, the probability that ball A was sufficient in this case is lower.

ROBUST-CAUSATION. Both philosophers (Lewis, 1986; Woodward, 2006) and psychologists (Lombrozo, 2010; Vasilyeva, Blanchard, & Lombrozo, 2018) have argued that another important aspect of causal relationships is their robustness. Causal relationships are robust to the extent that they would have continued to hold even if the background

conditions had been somewhat different.

Similar to how we defined sufficient-causation, we define robust-causation as

$$P_R(C \rightarrow e) = P_W(C \rightarrow e | change(\backslash C)). \tag{6}$$

After having observed what actually happened S, we consider a counterfactual situation in which all other candidate causes had been randomly perturbed ($change(\backslash C)$), and check whether in this situation, C would have been a whether-cause of the outcome. Remember that when testing for sufficient-causation, we consider a counterfactual situation in which the alternative causes had been removed from the scene. To test for robust-causation, we consider a counterfactual situation in which the alternative causes are still present but somewhat changed. For the specific case of colliding billiard balls, we can think of the *change*() operation as slightly perturbing the initial spatial location of the alternative causes.

The more certain we are that C would have been an whether-cause in a situation in which the alternative causes had been perturbed, the more robustly C brought about e. Robustness helps to differentiate between cases in which a candidate cause directly brought about an outcome (like ball A in Figure 3c), from situations in which the causal relationship was mediated by other candidate causes (like in the causal chain in Figure 3a).

Putting it all together

The CSM predicts that people's causal judgments are influenced by different counterfactual contrasts that determine the subjective degree of belief that the candidate was a whether-cause, a how-cause, a sufficient-cause, and a robust-cause of the effect event of interest.

Now that we have all the pieces that make up the CSM, let us say a little bit more about how to put them together. The CSM predicts that each of the different aspects of causation positively influences people's causal judgments. If participants believe that a candidate cause was a difference-maker, then their causal judgment is predicted to increase the more they believe that it was a whether-cause and a how-cause that was sufficient and robust. 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 important when judging causation. We will use the term "causal responsibility" to refer to the extent to which a candidate cause was considered to be "the cause" of the outcome. The overall causal responsibility of a cause C for an outcome event e is given by

$$\text{Causal responsibility}(C \rightarrow e) = P_{DM} \cdot (\beta_1 P_W + \beta_2 P_H + \beta_3 P_S + \beta_4 P_R). \tag{7}$$

TG: does this capture the right if/then switch structure? we don't want to model this as an interaction of DM with all the other aspects

Figure 6 illustrates the sequential nature in which the different counterfactual contrasts are considered. The model begins with a causal connection phase that selects only candidates that made a difference for how the outcome came about. The difference-making test selects amongst the candidates, the ones that were "a cause" of the outcome. For each

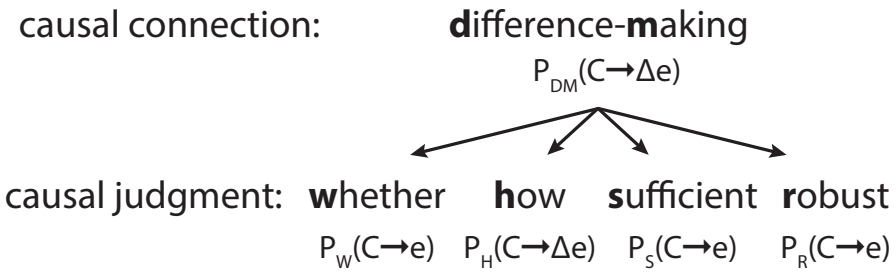


Figure 6. Relationship between the counterfactual contrasts. The counterfactual contrasts on the lower level are only considered if the test for difference-making is passed.

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.

Figure 7 shows graphically, how the different aspects of causation are evaluated by the model. The different tests have in common that they all define a counterfactual operation over the intuitive physical representation of the situation. The tests differ in terms of (1) what contingency they consider relevant, (2) the counterfactual contrast, and (3) the granularity at which the outcome event is specified.

In the example shown in Figure 7, the model evaluates the extent to which ball A caused ball E to go through the gate. When considering whether-causation and how-causation, the relevant contingency is the actual situation. For sufficient-causation, the relevant contingency is a situation in which all alternative candidate causes (here, ball B) are removed from the scene. For robust-causation, the relevant contingency is a situation in which the position of all alternative causes was slightly changed.

To determine whether the candidate cause was a whether-cause, sufficient-cause, or robust-cause, the outcome in the relevant contingency is contrasted with the outcome of a counterfactual situation in which the candidate cause was removed from the scene. To test for how-causation, instead of removing the candidate cause from the scene, the model generates a counterfactual contrast by applying a small perturbation to the candidate cause (here illustrated via a small change to the initial position).

When comparing the outcome event in the relevant contingency with the one in the counterfactual contrast, it is either construed coarsely, or finely. For whether-causation, sufficient-causation, and robust-causation, the model simply checks whether the target ball went through the gate, or missed the gate. For how-causation, the outcome event is construed finely and includes additional information about the time at which E went through the gate, and where exactly it went through (or missed).

Let us illustrate how the full model works based on the three example cases shown in Figure 3. The CSM defines each aspect of causation as a probability which expresses the observer's subjective degree of belief that the respective aspect of causation was true. For simplicity and ease of exposition, we will assume for the examples below that the observer has access to the ground truth. Hence, each aspect of causation is either true or false (see Table 1).

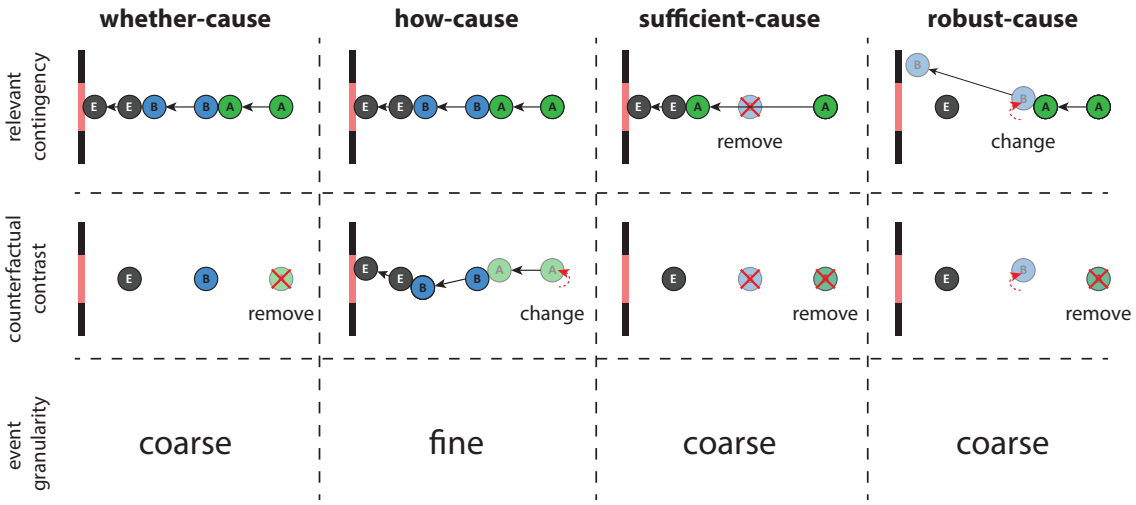


Figure 7. Illustration of the different types of counterfactual contrasts to determine the extent to which ball B caused ball E to go through the gate in the causal chain in Figure 3a. The top row shows the relevant contingency that serves as the starting point. This contingency is either the actual situation (for whether-causation and how-causation), or a situation in which the alternative cause (ball A) 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 the removal of the candidate cause (ball B), or a perturbation to it. The bottom row shows at what level of granularity the outcome events 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.

TG: maybe remove the lowest row on event granularity and just mention that in text

Causal chain. In the causal chain (Figure 3a), E and A are initially at rest. B enters the scene and collides with A. Ball A subsequently collides with E, and E goes through the middle of the gate. Both A and B are difference-makers in this situation. If either ball had been removed from the scene, then the outcome event (finely construed) would have been different from what it actually was. If B had been removed from the scene, then E would have just remained at rest. If A had been removed from the scene, then E would have gone through the gate slightly differently from how it actually did. It would have gone through the gate at a slightly earlier point in time assuming that there is some uncertainty about whether the collision was perfectly elastic. Since both balls were identified as difference-makers, the model proceeds to considering the other aspects of causation.

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

Ball B was a whether-cause. E would not have gone through the gate if B was removed from the scene. Ball B was also a how-cause. The outcome event would have been slightly different, if we had changed B's initial position. B was also a sufficient-cause. B's presence would have made a difference to whether or not E ended up going through the gate in a counterfactual situation in which ball A had been removed from the scene. Ball B was not a robust-cause of E's going through the gate. If A's initial position was slightly different from what it actually was, then removing B would have made a difference to whether or not E goes through the gate. If A's position was slightly changed (and B still entered the scene in exactly the same way), then E would not have gone through the gate because ball A would have missed it (see Figure 7). Removing ball B from the scene in that situation, would not have made a difference to the outcome.

TG: make sure that robustness and sufficiency are explained ok

Double prevention. In the double prevention case (Figure 3b), 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 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 E's going through the gate.

Ball A does not qualify as a difference-maker in this situation. If A had been reomved from the scene, then ball E would still have gone through the gate exactly in the same way as it did. Since ball A doesn't qualify as a difference-maker, the model does not consider any of the other aspects of causation for ball A.

Ball B was as a difference-maker. If it had been removed then A would have knocked E out of the way. Since B was a difference-maker, the model continues to consider the other aspects of causation to determine the extent to which B was responsible for E's going through the gate. B was a whether-cause. If B had been removed from the scene, then E would not have gone through the gate. B was not a how-cause. Even if B's initial position was perturbed, E would still have gone through the gate exactly in the same way that it did.[4] B was not a sufficient-cause. In a counterfactual situation in which the other candidate cause, ball A, was removed, B would not have made a difference to the outcome. If ball A had been absent, then E would gone through the gate no matter whether or not B had been present. B was also not a robust-cause of E's going through the gate. B's presence would not have made a difference to the outcome in a counterfactual situation in which ball A's

[4]There are of course ways in which we can change ball B such that E would not have gone through the gate. For example, if we change B's velocity sufficiently then it won't collide with A anymore. However, when testing for how-causation, we constrain ourselves to small changes. There are many small changes to ball B that would not make any difference to the spatiotemporal details of E's going through the gate in the double-prevention case. For the causal chain, in contrast, any change to any of the candidate cause balls affects the outcome event.

initial position was perturbed.

TG: ball B is in fact a robust difference-maker here; even if we apply a small perturbation to ball A, B is still a difference maker; update the paragraph

Preemption. In the preemption case (Figure 3c), ball E is initially at rest in front of the gate. Ball A collides with E and E goes through the gate. Ball B entered the scene in a way such that it would have knocked ball E into the gate just a moment later. Thus, ball A preempts ball B from knocking E into the gate.

Ball A was a difference-maker. The outcome event would have occurred differently if A had been removed from the scene. Ball A was not a whether-cause. E would have gone through the gate even if A had been removed from the scene (because of B). A was a how-cause. E would have gone through the gate differently, if A's initial position had been changed. A was also a sufficient-cause. It would have made a difference to whether or not E ended up going through the gate in a counterfactual situation in which B had been removed from the scene. Finally, A was also a relatively robust cause. There is a good chance that A's presence would have been pivotal in counterfactual situations in which B's initial position had been changed.

Ball B was not a difference-maker. 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. Again, since B did not qualify as a difference-maker, none of the other aspects of causation need to be considered.

General information about experiments

In the following, we will discuss the results of three sets of experiments that test how well the CSM captures participants' causal judgments. Before going into the specifics of each individual experiment, we will discuss the general aspects they have in common.

Description of the stimuli

All experiments were designed using Adobe Flash or Javascript. The videos were created using the flash and javascript implementation of the 2D physics engine box2d.[5] The basic setup for the stimuli in all experiments was identical. Participants viewed collision events between billiard balls from a bird's view perspective. Balls either entered the scene from the right or were present in the scene from the beginning and at rest. The scene was bounded by solid walls on the top, bottom, and left side. A small gate in the middle of the left-side wall was indicated by a red line. There was no friction and collision events were perfectly elastic, that is, there was no loss of momentum during collision events. 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 $13\frac{1}{3} \times 10\,\text{m}$ in the physics world). The radius of each ball was $0.5\,\text{m}$. In experiment 2, some of the clips also featured a static rectangular brick $\frac{1}{4} \times \frac{5}{6}\,\text{m}$ or a teleport with a yellow rectangle as entrance ($\frac{1}{4} \times \frac{5}{6}\,\text{m}$) and a blue circle as exit (radius $= \frac{1}{3}\,\text{m}$).

[5] See `http://box2d.org/` for the box2d engine, and `http://www.box2dflash.org/` as well as `https://github.com/hecht-software/box2dweb` for the flash and javascript implementations, respectively.

Table 1
Results of the different counterfactual tests applied to balls A and B for the situations depicted in Figures 3 and 5. Note*: If a ball doesn't qualify as a difference-maker, none of the subsequent tests are considered. The values in parentheses show the quantitative predictions of the CSM (which is described in detail in Section XX.)*

Situation	Ball	$P_{\text{DM}}(C \to \Delta e)$	$P_{\text{W}}(C \to e)$	$P_{\text{H}}(C \to \Delta e)$	$P_{\text{S}}(C \to e)$	$P_{\text{R}}(C \to e)$
causal chain	A	✓ (1)	✗ (0.05)	✓ (1)	✗ (0)	✗ (0.04)
	B	✓ (1)	✓ (1)	✓ (1)	✓ (1)	✓ (0.99)
double prevention	A	✗ (0)	– (0)	– (0)	– (0)	– (0)
	B	✓ (1)	✓ (0.99)	✗ (0)	✗ (0.23)	✓ (0.63)
preemption	A	✓ (1)	✗ (0.27)	✓ (1)	✓ (0.99)	✗ (0.24)
	B	✗ (0)	– (0)	– (0)	– (0)	– (0)
joint causation	A	✓ (1)	✓ (0.8)	✓ (1)	✗ (0.41)	✓ (0.76)
	B	✓ (1)	✓ (0.83)	✓ (1)	✗ (0.18)	✓ (0.84)
overdetermination	A	✓ (1)	✗ (0)	✓ (1)	✓ (1)	✗ (0)
	B	✓ (1)	✗ (0)	✓ (1)	✓ (1)	✗ (0)

Note: DM = difference-making, W = whether-cause, H = how-cause, S = sufficient-cause, R = robust-cause.

TG: double check that the examples and table match

TG: update link to the model section

For Experiments 1 and 2, all clips featured a single collision event between balls A and B. The clips in Experiment 3 featured three balls, and the number of collision events varied between clips.[6]

Experimental design

The different clips in each experiment varied in terms of what happened in the actual situation as well as what would have happened if a ball had been removed from the scene. Each experiment was run in two conditions: a counterfactual and a causal condition. In the *counterfactual judgment condition*, participants were asked to judge whether they thought the target ball would have gone through the gate if the candidate cause ball had been absent from the scene. In the *causal judgment condition*, participants were asked to evaluate the causal role of the ball(s) of interest. For experiments that only featured a single candidate cause, we asked participants to judge what role ball A played – that is, whether it caused B to go through the gate, or prevented B from going through. For the experiment that featured two candidate causes, we asked participants to judge to what extent each ball was responsible for E's going through the gate, or E's not going through the gate. All judgments were elicited on sliding scales that allowed participants to make graded judgments.

[6]The complete materials including the video clips, diagrams, and data may be accessed here: `https://github.com/tobiasgerstenberg/csm`

Experimental procedure

All 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. Participants were only allowed to participate in a single experiment. No participants were excluded from any of the experiments.

Experiment 1: Simple collision events

In Experiment 1 we test the predictions of the *counterfactual simulation model* (CSM) for clips that involve a single candidate cause. Participants saw 18 different clips which varied whether ball B clearly missed the gate 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. B would have either clearly missed the gate ("counterfactul miss"), just missed or barely gone through the gate ("counterfactual close"), or clearly gone through the gate ("counterfactual hit"). Crossing actual with counterfactual closeness generates nine qualitatively different sets of stimuli. For each of the nine combinations, we created two different clips. Ball B actually went through the gate in half of the clips and it also would have gone through the gate if ball A hadn't been present in the scene in half of the clips. 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. Figure 8 shows diagrams of the different clips. The solid arrows in indicate the actual paths that balls A and B moved on before they collided, as well as the path that ball B moved on after the collision. The dashed arrows indicate how ball B would have moved if ball A hadn't been present in the scene.

For example, in clip 1, ball B clearly missed the gate. It would also clearly have missed if ball A had been absent from the scene. In clip 9, B just missed the gate. If A had been absent from the scene it would have barely gone in. In clip 17, ball B went through the middle of the gate. It's also clear that B would have gone through the gate if ball A had been removed from the scene.

What question participants were asked was varied between conditions. One group of participants was asked to make counterfactual judgments. The other group of participants was asked to make causal judgments. Participants in both conditions saw each clip twice before making their judgment. We will discuss the results of each condition in turn.

Counterfactual judgments

In the *counterfactual judgment condition* the clips were paused shortly after the time at which the balls collided. Upon having viewed the clip for a second time, the following question probe 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 that ranged 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.

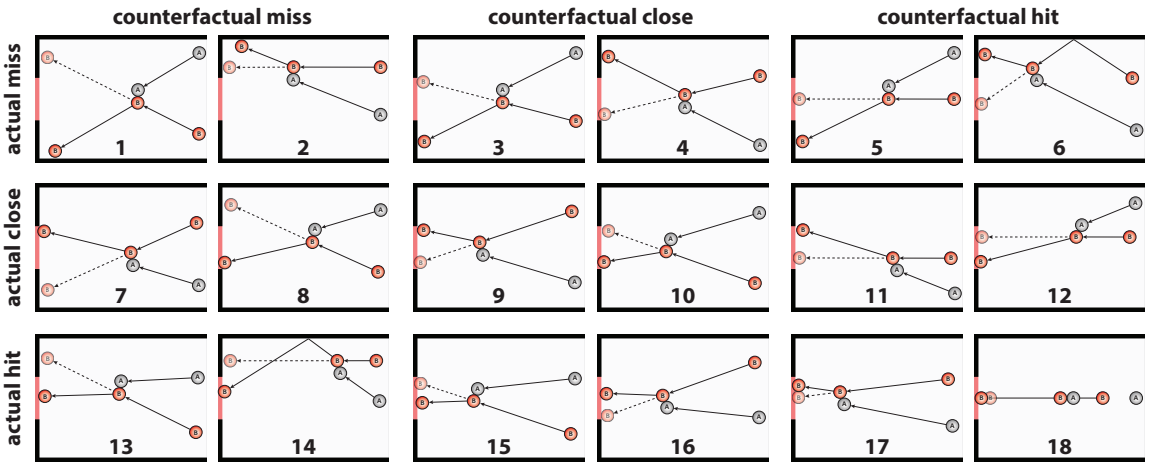


Figure 8. Diagrams of clips used in Experiment 1. *Note*: The solid arrows indicate the actual paths on which balls A and B moved before the collision and the path that ball B moved after the collision. The dashed line indicates the path on which ball B would have moved if ball A had not been present in the scene. The numbers indicate the clip number.

TG: check whether it's a problem that this figure was used before

Participants and Procedure. 41 participants (19 female, $M = 32$ years, SD = 8.76) participated in the experiment. Participants were instructed that they would see 18 different video clips, and that their task would be to make counterfactual judgments. The order of the clips was randomized. On average, it took participants 7.5 minutes (SD = 3.12) to complete the experiment.

Design. The experiment followed a 3 (*actual outcome closeness*: clear hit, close hit, clear miss) × 3 (*counterfactual outcome closeness*: clear hit, close hit, clear miss) design, whereby participants saw two different clips for each combination of actual and counterfactual outcome closeness (cf. Figure 8).

Results and Discussion. Figure 9 shows participants' mean counterfactual judgments together with the predictions of the best-fitting approximate simulation model (cf. Section **??**"Modeling counterfactual simulations"). Recall that the approximate simulation model introduces some noise into the simulation of what would have happened if ball A had been removed from the scene. This noise is introduced in the form of a random perturbation to the direction of ball B's velocity vector at each time step in the simulation after the point at which the two balls would have collided. In order to get the model predictions, we generated 1000 noisy samples for each of the 18 clips using different degrees of noise ranging from SD = 0° to SD = 2° in steps of 0.1°, where SD refers to the standard deviation of the Gaussian distribution from which the perturbation to B's velocity vector was drawn.

TG: make an interface for which people can test the degree of noise

The black bars in Figure 9 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 1, ball B ended up going through the gate only in 5 out of 1000 cases. In clip 3, B went through the gate in 447 out of 1000 cases, and in clip 5, it went through the gate in the 945 out of

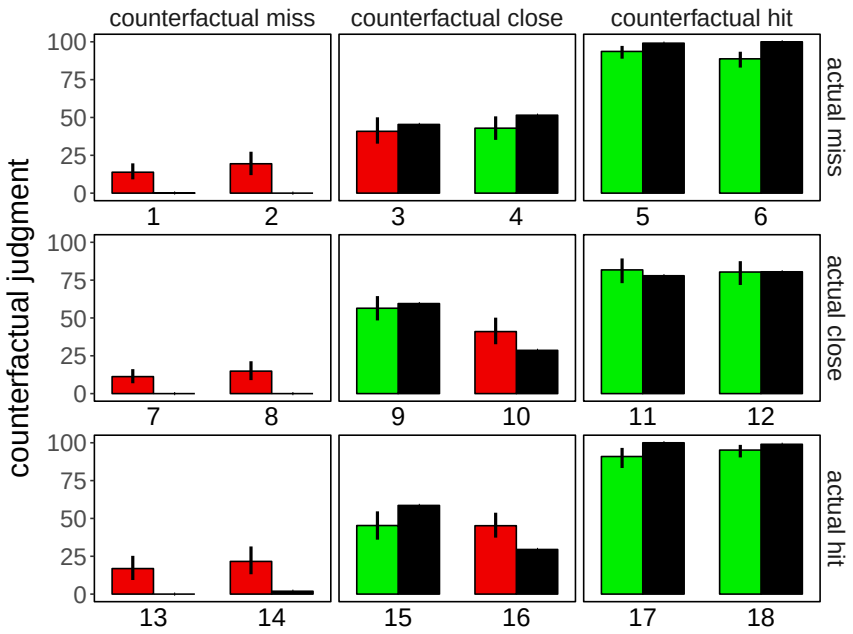


Figure 9. Mean counterfactual judgments (colored bars) together with the model predictions (black bars) of the best-fitting approximate simulation model. *Note*: Red bars indicate cases in which ball B would have missed. Green bars indicate cases in which B would have gone through the gate. 0 = "ball B would have definitely missed", 100 = "ball B would have definitely gone through the gate". Error bars indicate bootstrapped 95% confidence intervals.

TG: add additional indiactor to red bars for grayscale version

1000 cases. We chose the noise parameter that maximizes the correlation between model prediction and participants' mean counterfactual judgments. The approximate simulation model captures people's counterfactual judgments best at a noise level of 0.5° with $r = .979$ and RMSE = 11.866. For comparison, a deterministic physics model which applies no noise explains people's counterfactual judgments less well with $r = .833$ and RMSE = 29.999. This model simply predicts a rating of 0 for the cases in which ball B would have missed (the red bars in Figure 9) and a rating of 100 for the cases in which ball B would have gone in (the green bars).

As Figure 9 shows, the model slightly overestimates participants' certainty for cases in which ball B clearly misses (first column). The model also predicts that participants assign a greater chance that B would have gone through the gate in clip 15 compared to clip 16. However, participants' judgments were almost identical for these cases. Overall, however, there is a very close fit between the model's predictions and participants' counterfactual judgments.

The results of the counterfactual condition show that people are capable of simulating whether B would have gone through the gate if ball A had been removed from the scene. By assuming that people's mental simulations of where ball B would have gone if ball A

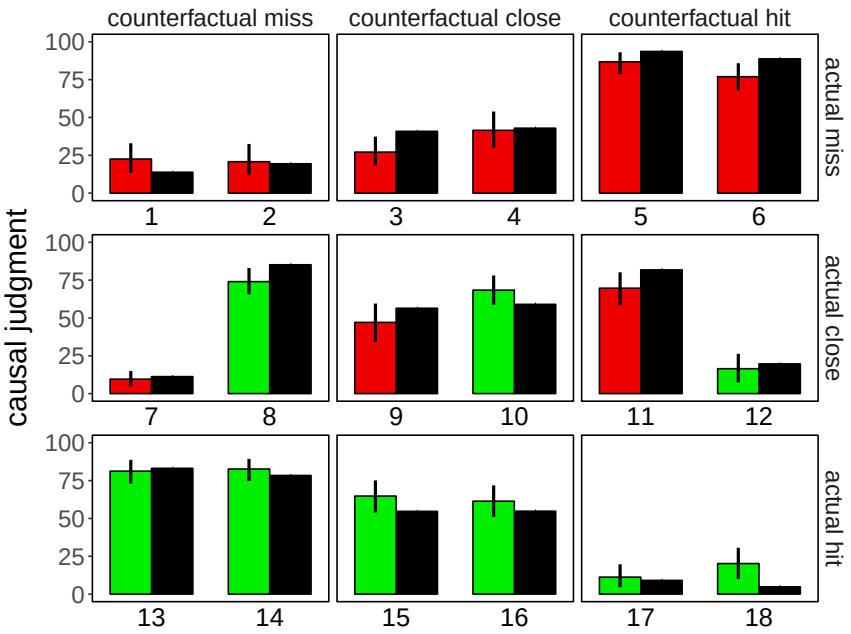


Figure 10. Causal judgments with model predictions. *Note*: Red bars are prevention judgments and green bars are causation judgments. Error bars indicate bootstrapped 95% confidence intervals.

hadn't been present are somewhat noisy, we can accurately capture people's uncertainty in the counterfactual outcome.

Causal judgments

In the *causal judgment condition*, participants saw each clip played twice until the end. They were then asked on a separate screen to to answer the question: "What role did ball A play?" Participants indicated their responses 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 A somewhat caused 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.

Participants and Procedure. 41 participants (33 female, $M = 33$ years, SD = 10.3) participated in the experiment. Participants were instructed that they will see 18 different video clips and that their task will be to make causal judgments. The order of the clips was randomized. On average, it took participants 9.65 minutes (SD = 6.58) to complete the experiment.

Design. The design was identical to the counterfactual judgment condition.

Results. Figure 10 shows participants' average causal judgments together with the predictions of the CSM. We reverse coded the prevention ratings so that both cause and prevention ratings are on a scale from 0 to 100. Note that all the clips in this experiment only involved one candidate cause that directly interacted with the target ball. Since the

two balls collide in all clips, ball A trivially passes the test for difference-making in all of the situations (cf. Equation 2). The fact that both balls had direct contact also implies that A was a how-cause of the outcome event in all situations (cf. Equation 4). If participants only cared about whether A was a how-cause, then we would expect maximal ratings for each situation. Further, since there were no other candidate causes, tests for sufficiency (Equation 5) and robustness (Equation 6) are passed trivially. This means that for the clips in this experiment, the CSM reduces to simply considering if ball A was a whether-cause of B's going through the gate, or missing the gate.

As discussed in Section **??**"whether-causation", we have two ways of determining the probability that A was a whether-cause. We can either take the predictions of our approximate simulation model, or directly use participants' counterfactual judgments. Here – and for all of the other experiments – we do the latter. Recall that the probability that A was a whether-cause equals the probability that the outcome would have been different from what it actually was if ball A had been removed from the scene (cf. Equation 3).

In situations in which B went through the gate, we need the probability that B would *not* have gone through the gate if ball A had been absent. When B missed the gate, we need the probability that B would have gone through the gate in A's absence. Since we asked participants in the counterfactual judgment condition whether 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 B would *not* have gone through the gate in the absence of A.

As Figure 10 shows, the CSM's model predictions and participants' causal judgments are very closely aligned. To determine the probability P_W that A was a whether-cause of B's going through the gate (or missing the gate), we directly use participants' mean counterfactual judgments without applying any transformation. The correlation between model and data is very high with $r = .961$ and RMSE = 8.57. The CSM's fit is also high when we use the predictions of the best-fitting approximate simulation model to determine P_W, with $r = .946$ and RMSE = 16.29.

As the results show, participant's cause and prevention judgments were determined by their subjective degree of belief that ball A made a difference to whether or not B went through the gate. Participants' causal judgments (the green bars in Figure 10) were highest when it was clear that ball B would have missed if ball A hadn't been present in the scene (clips 8, 13, and 14). Their causal judgments were intermediate for situations in which it was unclear whether B might have gone through the gate even if ball A hadn't been there (clips 10, 15, and 16). Their judgments were lowest for situations in which ball B would have gone in anyhow even if ball A hadn't been present (clips 12, 17, and 18).

Similarly, for prevention judgments, participants' judgments were highest in situations in which it was clear that ball B would have gone if ball A had been absent (clips 5, 6, and 11). Their prevention judgments were intermediate in situations in which it was difficult to tell whether B would have missed the gate if ball A hadn't been there (clips 3, 4, and 9). Finally, their prevention judgments were lowest in situations in which it was clear that B would have missed even if ball A hadn't been present in the scene (clips 1, 2, and 7).

Interestingly, participants' judgments were not affected by how closely B ended up going through the gate, or ended up missing the gate. For example, in clip 8 ball B

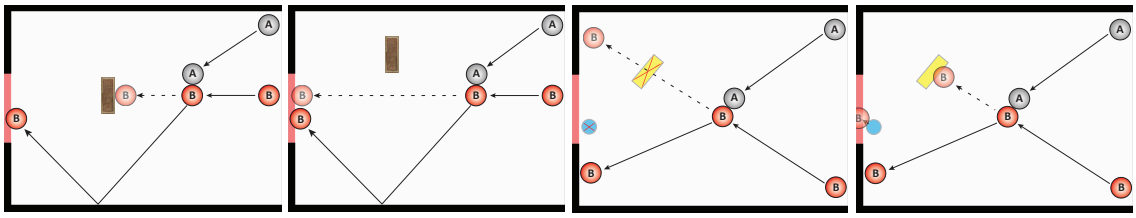
just barely goes through the gate. It was clear that it would have missed but for A. In clip 13, in contrast, ball B goes right through the middle of the gate (and, again, it's clear that it would have missed but for A). Participants' judgments were almost identical in both cases with $M = 74$ (SD $= 28.88$) for clip 8 and $M = 78.66$ (SD $= 33.29$) for clip 13, $t(40) = -0.825, p = .414, d = 0.129$. More generally, holding the outcome fixed, participants' judgments did not vary as a function of actual closeness of the outcome (the rows in Figure 10, with $F(1, 40) = 1.86, p = 0.18, \eta^2 = 0.005$ for negative outcomes, and $F(1, 40) = 0.01, p = 0.935, \eta^2 = 0$ for positive outcomes) but did vary strongly as a function of the counterfactual closeness of the outcome (the columns in Figure 10, with $F(2, 80) = 65.23, p = 0, \eta^2 = 0.369$ for negative outcomes, and $F(2, 80) = 96.33, p = 0, \eta^2 = 0.465$ for positive outcomes).

Discussion

The results of Experiment 1 show that the CSM accounts well for people's causal judgments in a physical domain that features collisions between billiard balls. People's counterfactual judgments about whether one of the balls would have gone through the gate if the other ball had been absent, are well-accounted for by an approximate simulation model that assumes that people's mental simulations of what would have happened in the relevant counterfactual situation are somewhat noisy (cf. Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2013). A version of the approximate simulation model which introduces small random perturbations to the target ball's direction of velocity after the point at which the two balls would have collided accurately captures people's subjective degree of belief about what would have happened.

The results of the causal judgment condition show that people's judgments are very closely in line with the predictions of the CSM. Since all of the clips featured direct contact between the balls, the candidate ball was a how-cause in all of the clips. Since only a single candidate cause was present, the sufficiency and robustness remained constant throughout the clips. Thus, the only aspect of causation that is required to explain participants' causal judgments in this experiment is whether-causation which captures the extent to which the presence of the candidate cause was necessary for bringing about the outcome.

While the results of Experiment 1 provide good support for the role of counterfactual simulation in people's causal judgments, they do not rule out alternative explanations for people's judgments. As apparent from Figure 8, the set of clips in the experiment 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 is possible in principle to provide an account of the results by appealing to differences merely in what actually happened and without the need to rely on counterfactuals. We doubt that an adequate account of the results in this experiment can be developed that does not rely on counterfactual simulation. However, rather than speculating whether an actualist account may explain people's causal judgments, we will see in Experiment 2 that causal judgments are inextricably linked to counterfactual simulation.



(a) Ball B would have been blocked by the brick

(b) Ball B would have gone through the gate

(c) Ball B would have missed

(d) Ball B would have gone through the gate via the teleport

Figure 11. Illustration of two pairs of clips in which the interaction between balls A and B is identical, but the relevant counterfactuals are different. For each pair, there is one clip for which the outcome (broadly construed) would have been different if ball A hadn't been present in the scene (a and c), and one clip for which the outcome would have been the same (b and d).

Experiment 2: Bricks and Teleports

Experiment 2 has two goals: First, to provide even stronger evidence for the use of counterfactual simulation in people's causal judgments. Second, to demonstrate the flexibility of people's mental simulations. In Experiment 1, the world was relatively simple – it only featured two billiard balls, some walls, and a gate. In this experiment, the world is a little more complex. Some of the clips now feature a solid brick that is placed somewhere in the scene. Other clips feature a teleport that, if active, transports ball B from one location to another.

In this experiment, the set of clips included pairs for which what actually happened was held constant but what would have happened if ball A had not been present in the the scene was different. For example, consider the pair of clips shown in Figure 11a and 11b. 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 11a, ball B would not have gone through the gate even if A hadn't been present in the scene because it would have been blocked by the brick. In Figure 11b, in contrast, the brick's location is different. Here, 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 in terms of what actually happened (which was identical in both clips).

When we use the the brick to manipulate what would have happened in the relevant counterfactual, we need to change what the scene actually looks like. We also generated a number of clips in which we manipulated the counterfactual without changing the spatial location of any components in the scene. For this purpose, we introduced participants to the teleport. The teleport only affects ball B but not ball A. It has an entry (the yellow rectangle) and an exit (the blue circle). If ball B enters the teleport through the yellow rectangle, it leaves the teleport through the blue circle in the same direction in which it entered the teleport.

Consider the pair of clips shown in Figure 11c and 11d. Again, in both clips, what actually happened was identical. In Figure 11c, the teleport was switched off (as indicated

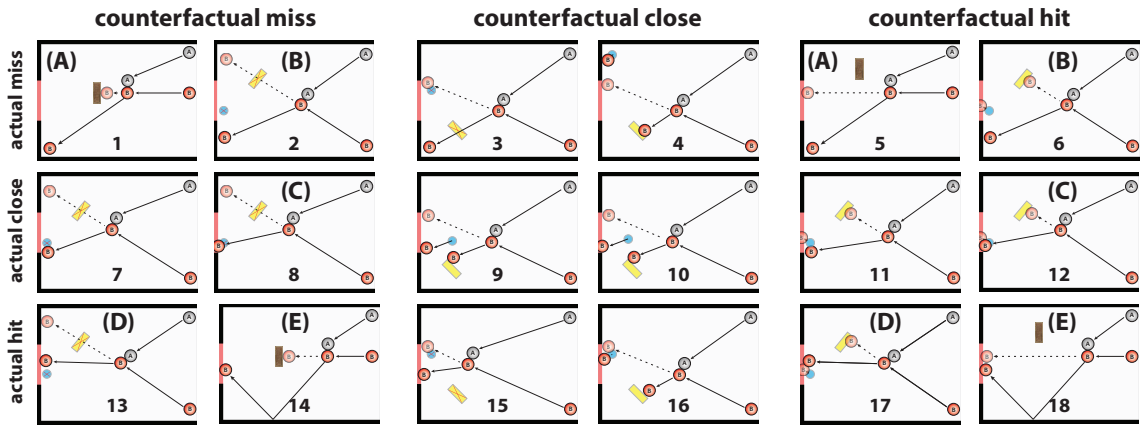


Figure 12. Diagrams of clips used in Experiment 2. *Note*: The solid arrows indicate the actual paths on which balls A and B moved before the collision and the path that ball B moved after the collision. The dashed line indicates the path on which ball B would have moved if ball A had not been present 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.

by the red crosses on top of the teleport entry and exit). Hence, B would have missed the gate even if ball A hadn't been present in the scene. In contrast, in Figure 11d, the teleport was on. Here, B would have gone through the gate (via the teleport), if ball A had not been present in the scene.

TG: be consistent about saying "removed" vs. "had not been present"

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. In experiment 1, the counterfactual simulations participants needed to do in order to evaluate whether A made a difference to whether B went through the gate involved relatively simple extrapolations of B's movement. In this experiment, some of the counterfactual simulations are more challenging in that they involve taking into account the way in which ball B interacts with the teleport. Successful mental simulation in these cases would demonstrate that participants can go beyond simple physical interactions to infer what would have happened.

Methods

The basic setup of this experiment replicates the setup of Experiment 1. Figure 12 shows diagrams of the clips that participants viewed in this experiment. Again, we systematically varied the closeness of the actual outcome (rows) as well as the closeness of the counterfactual outcome (columns). The letters (A) through (E) indicate pairs of clips in which what actually happened was identical, but the outcome in the relevant counterfactual would have been different.

Participants and Procedure. 82 participants (45 female, $M = 34$ years, SD = 12.2) participated in the experiment. Participants were instructed that they will see 20 different video clips in total (which included two practice clips). In this experiment, we had each participant provide both counterfactual judgments as well as causal judgments. Half of the participants first made counterfactual judgments and then causal judgments. For the other half of participants, the order of blocks was reversed. Participants always saw two practice clips first. The practice clips served to familiarize participants with the block as well as how the teleport worked. The order of the 18 test clips was randomized in each of the two judgment blocks. On average, the experiment took $M = 21.2$ (SD = 5.11) to complete.

Both the causal and counterfactual judgments were elicited exactly in the same way as in Experiment 1. At the beginning of the experiment, participants did not know that they will be asked to make both counterfactual and causal judgments. That is, 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) × 3 (*counterfactual outcome closeness*: clear hit, close hit, clear miss) × (*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 12. In contrast to Experiment 1 which varied the type of judgment between participants, all factors were varied within participants.

Results and Discussion

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

Counterfactual judgments. Because there was neither a main effect of question order, $F(1, 80) = 0.47, p = 0.495, \eta^2 = 0$, nor an interaction effect of question order and clip number, $F(17, 1360) = 0.33, p = 0.996, \eta^2 = 0$, we will combine the counterfactual judgments for both question orders. Figure 13 shows participants' mean counterfactual judgments for the different clips together with the predictions of the best-fitting approximate simulation model. The best-fitting approximate simulation model has a noise value of 0.6° with $r = .967$ and RMSE = 14.41. For comparison, a deterministic model (i.e. 0°) again performs worse with $r = .864$ and RMSE = 28.14. The results of this experiment show that participants have no trouble simulating counterfactuals in more complex situations that require taking into account the operation of a teleport. For example, in clip 17, participants are almost as certain as they are in clip 18 that ball B would have gone through the gate if ball A had been absent from the scene. Having established that participants can assess what would have happened if ball A had been removed from the scene, we will now look at how participants made causal judgments in this setup.

Causal judgments. There was again no main effect of question order on participants' causal judgments, $F(1, 80) = 0.70, p = 0.405, \eta^2 = 0$. However, there was an interaction effect between question order and clip number, $F(17, 1360) = 3.20, p < .001, \eta^2 = 0.04$. Figure 14 shows the causal judgments separated by whether participants first made counterfactual judgments (Figure 14a) or causal judgments (Figure 14b). We again reverse coded

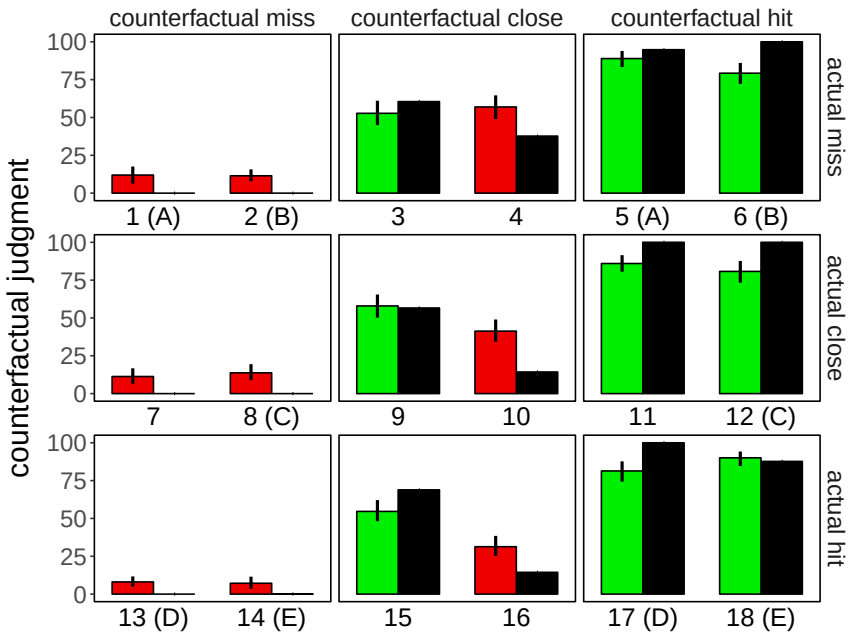


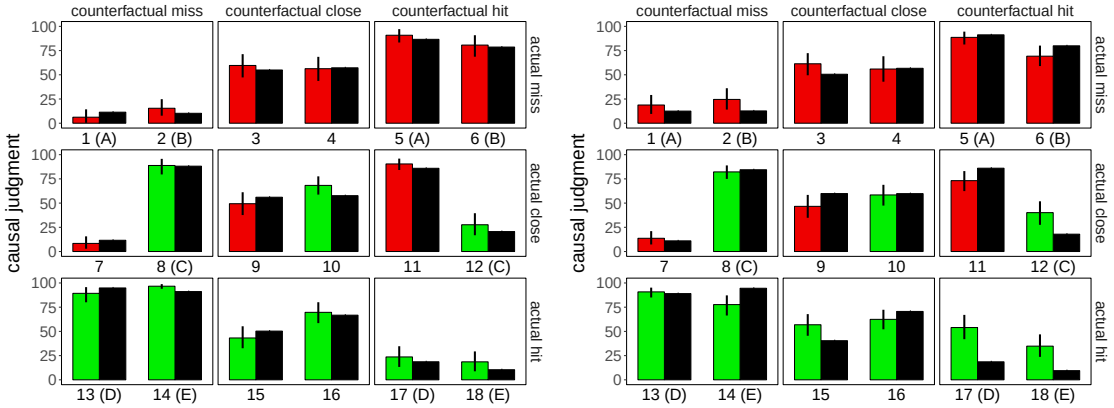
Figure 13. Mean counterfactual judgments (colored bars) together with the model predictions (black bars) of the best-fitting approximate simulation model. *Note*: Red bars indicate cases in which ball B would have missed. Green bars indicate cases in which B would have gone through the gate. Error bars indicate bootstrapped 95% confidence intervals.

the prevention ratings so that both cause and prevention ratings are on a scale from 0 to 100.

Let us first focus on the commonalities between both conditions before talking about the differences. Like in Experiment 1, the only aspect of causation that was manipulated between the clips was the probability that ball A was a whether-cause of B's going through the gate (see Equation 3). Since both balls directly collided in all of the clips, A was a how-cause in all of the clips, and because A was the only candidate cause, A was a sufficient and robustness cause in all situations. Again, we use participants' counterfactual judgments to determine P_W for each clip. For those participants who answered the counterfactual questions first, the correlation between their counterfactual and causal judgments was very high with $r = .985$ and RMSE $= 5.52$. For the other group of participants who made causal judgments in the first block and counterfactual judgments in the second block, the correlation was lower but still high with $r = .911$ and RMSE $= 14.51$.

Just like in Experiment 1, 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 B went through the gate (see Equation 3). 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 14a and 14b. Again, how close ball B actually ended up going through the gate had no effect on participants' judgments (compare different rows in Figures 14a and 14b).

TG: add stats here?



(a) Counterfactual judgments first (b) Causal judgments first

Figure 14. 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. Error bars indicate bootstrapped 95% confidence intervals. Error bars indicate bootstrapped 95% confidence intervals.

Moreover, 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 (as marked by letters (A) to (E)). For example, for clip 1, participants did not think that ball A prevented ball B from going through the gate. In this situation, the brick would have blocked B even if ball A hadn't been present in the scene. In contrast, for clip 5, participants judged that A prevented B from going through the gate. Here, the brick wouldn't have blocked B. Thus, even though the way in which balls A and B move and collide is 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 off (clip 2) or on (clip 6). 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 conditions, there were also some differences. Participants' judgments between conditions differed most strongly for clips 12, 17, and 18. Whereas participants who made counterfactual judgments first, gave very low causal ratings for these clips, participants who made causal judgments first, gave much higher ratings.

Discussion

The results of Experiment 2 show that causal judgments are fundamentally linked to counterfactual considerations. Participants' causal judgments differed significantly between clips that were matched in terms of what actually happened and only differed in what would

have happened if the candidate cause had been removed from the scene. This demonstrates that counterfactual simulation is necessary for causal judgment and that it won't be possible to develop an actualist account that adequately captures people's causal judgments.

Despite the fact that the setup in Experiment 2 was richer than in Experiment 1 in that it featured a brick as well as a teleport, participants had no trouble simulating what would have happened if ball A 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 can capture people's counterfactual judgments very accurately. Experiment 2 further demonstrates that people's mental simulations are very flexible. Even though we don't normally encounter teleports in our everyday life, participants had no trouble simulating what would have happened in counterfactual situations that included the operation of the teleport.

Like in Experiment 1, Experiment 2 showed that there is a tight relationship between their causal judgments and and their subjective degree of belief about whether the candidate cause made a difference to whether or not the outcome (broadly construed) occurred. The more certain people are that the candidate object was a whether-cause, the higher the causal rating. Participants made intermediate causal judgments in situations where they were unsure about whether the outcome would have changed had the candidate cause been removed from the scene. Again, the actual closeness of the outcome did not affect participants' causal judgments.

While in Experiment 1, we had varied the type of question (causal vs. counterfactual) between participants, in Experiment 2 we asked each participant both questions in different blocks. The results revealed an 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 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. When these participants were then later asked to make causal judgments, it could well be that they remembered what would have happened in the relevant counterfactual situation for this clip. 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. However, this explanation would suggest that the differences between conditions should be stronger in situations in which the counterfactual outcome was unclear (i.e. the middle columns in Figure 14) than in situation in which the counterfactual outcome was clear.

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 in situations in which A's presence made no difference to whether 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 causal than for prevention judgments.

Experiment 3: Complex causal interactions

Experiments 1 and 2 looked at situations that featured a single candidate cause. The results show that participants' causal judgments are strongly influenced by the extent to which the candidate cause was perceived as a whether-cause of the outcome. However, the clips in these experiments did not manipulate the other aspects of causation that the CSM postulates. Experiment 3 tests the CSM more comprehensively. By looking at situations that involve two candidate causes, we can tease apart the different aspects of causation and see how they affect participants' causal judgments.

Methods

The general setup in Experiment 3 was identical to the one used before. However, this time, the clips included two candidate causes, balls A and B, and one candidate target, ball E. Figure 15 shows diagrams of the 32 different clips that participants viewed in this experiment. Remember that for Experiments 1 and 2, 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 not been present in the scene. Given that there are four relevant 'worlds' (actual, only ball A, only ball B, neither ball A nor ball B) for which 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 15 shows a case in which E actually did not go through the gate. E would also not have gone through the gate if either ball B or ball A had been removed from the scene. Finally, ball E would also not have gone through the gate if both ball A and B had been removed. For clip 23, in contrast, ball E actually 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. As discussed in Section "The Counterfactual Simulation Model", 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, we manipulated between participants whether participants were asked to make counterfactual, or causal judgments. We discuss the results from both conditions in turn.

Counterfactual judgments

The CSM captures different aspects of causation in terms of different counterfactual operations (see Figure 7). For example, in order to determine the extent to which a can-

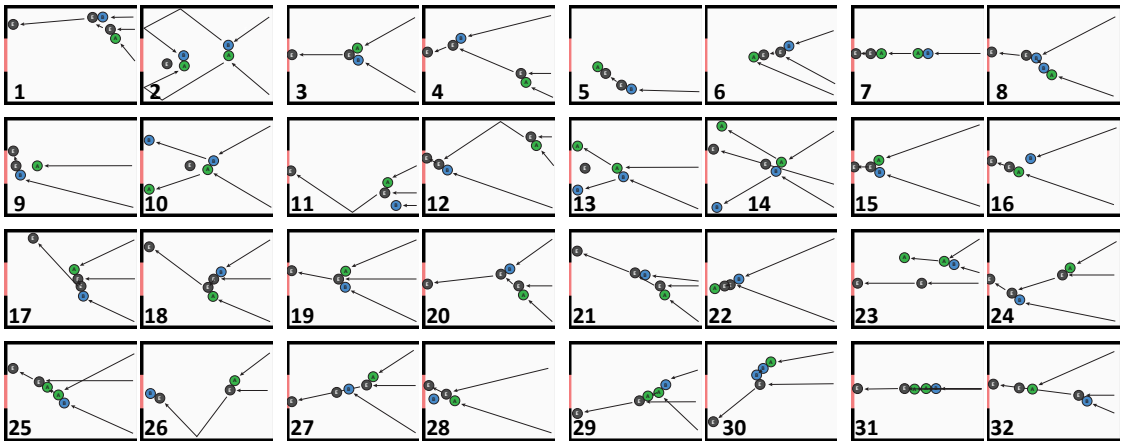


Figure 15. Diagrams of the clips used in Experiment 3. The clips varied whether ball E went through the gate in the actual situation, and what would have happened if either ball A, ball B, or both balls had been removed from the scene. See Table A1 for more information about each clip.

didate cause was a whether-cause of the outcome, the model removes the cause from the actual situation and estimates how likely the outcome would have been different.

In Experiments 1 and 2, this counterfactual simulation only required extrapolating where ball B would have gone if ball A hadn't been present in 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 15), we need to simulate whether ball A would have collided with ball E, and whether this collision would have lead to ball E missing the gate. Here, we look at whether participants are capable of simulating what would have happened in these more complex situations.

Participants and procedure. 80 participants (34 female, M = 33.4 years, SD = 10.1) participated in the experiment. Half of the participants made counterfactual judgments about ball A, and the other half about ball B. Participants were instructed that they will 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/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. This was done to remind participants of what the actual clip had looked like. On average, it took participants 18.1 (SD = 4.63) minutes to complete the experiment.

Results. Figure 16 shows participants' mean counterfactual judgments together with the predictions of the approximate simulation model. To predict participants' coun-

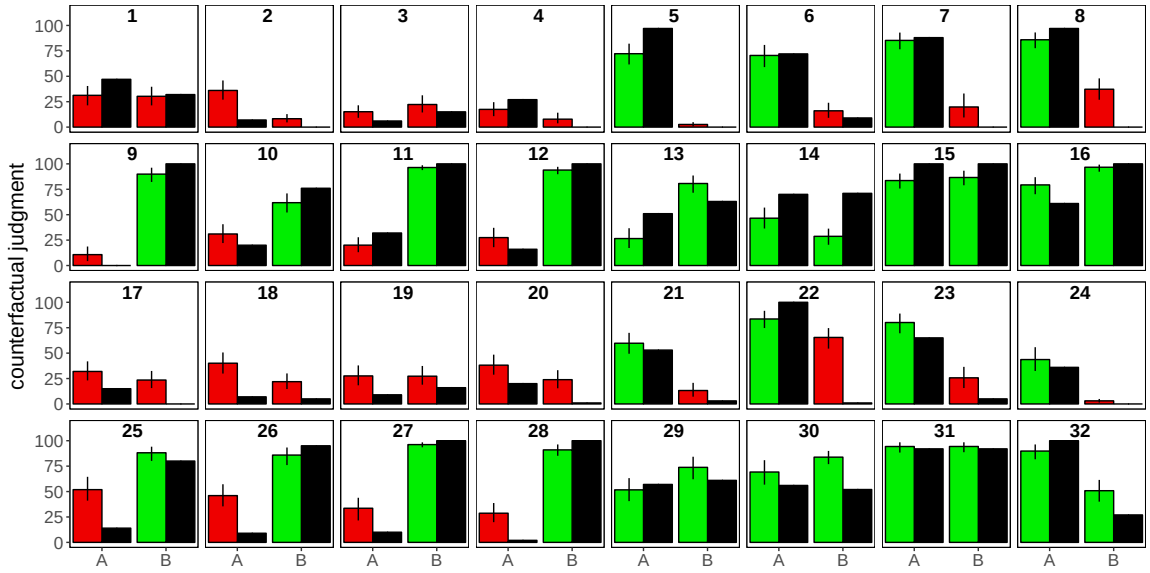


Figure 16. Mean counterfactual judgments (colored bars) together with the model predictions (black bars) of the best-fitting approximate simulation model. For each clip, the two bars indicate participants' belief that ball E would have gone through the gate if ball A hadn't been present in the scene (left bar) or if ball B hadn't been present (right bar). For example, in clip 11, participants believed that E would not have gone through the gate if A hadn't been present, but would have still gone through the gate if B hadn't been present. *Note*: Red bars indicate cases in which ball E would have missed the gate. Green bars indicate cases in which E would have gone through. Error bars indicate bootstrapped 95% confidence intervals.

terfactual 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. This noise is added at each step in time, beginning at the point at which the removed ball would have collided first with any of the remaining balls. Overall, the approximate simulation model fits participants' counterfactual judgments well with $r = .88$ and RMSE $= 19.40$ using a noise parameter of SD $= 1.3°$. A deterministic physics model (i.e. SD $= 0°$) does worse with $r = .829$, RMSE $= 30.28$.

Discussion. Overall, participants' counterfactual judgments were again well described by the approximate simulation model even for these more complex clips for which the relevant counterfactual simulation involved considering how multiple balls would move. Note that the degree of noise that was required to best account for participants' judgments (SD $= 1.3°$) is somewhat higher than it was in Experiment 1 (SD $= 0.5°$) and Experiment 2 (SD $= 0.6°$). This is likely due to the fact that simulating the relevant counterfactuals in this experiment often involves computing the outcome of a collision between the two remaining balls. For example, in clip 10, 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. Since Experiments 1 and 2 only featured two balls, the relevant counterfactual was easier to compute and never involved a collision between two balls.

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

TG: double check these parameters

TG: last point might be put a little more succinctly; maybe add a reference to the

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 (see, for example, the judgments for ball A in clips 25–28). When evaluating whether E would have gone through the gate if ball B hadn't been present in clip 22, participants have to gauge whether E would have managed to pass by ball A which was lying in front of the gate. Participants considered it likely that E would have gone through the gate if B had been removed, while in fact it would not have. Indeed, it would have required a significant amount of noise (in the right direction) for E to pass by A, and the model thus predicts that participants' judgment in this case should be low. Overall, however, the model does a fine job at capturing participants' beliefs about what would have happened.

TG: maybe mention clip 14 as well? it's tricky here to compute the collision and people are less likely than the model to believe that ball E would have gone through the gate

Causal responsibility judgments

Based on participants' judgments in the counterfactual condition just discussed, we can determine the extent to which each candidate cause qualifies as a whether-cause of E's going through the gate (cf. Figure 7). To be able to apply the full counterfactual simulation model (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 (cf. Figure 6). We will first discuss in detail how the CSM was implemented, and then look at how well it accounts for participants judgments.

Model prediction.

TG: discuss one model that uses participants' judgments for whether-causation, and one model that uses the model predictions?

TG:

- discuss the different versions of the tests → local noise vs. global noise?
- maybe mention here that the number of simulations (2 for each test) is plausible given what we know about the eye-tracking data?
- whether-causation:
  - global noise model: $r = .88$
  - local noise model: $r = .87$

Recall that the CSM first uses a test for difference-making to establish whether a candidate cause was causally connected to the outcome. Only if this connection was established, are the remaining aspects of causation considered to determine the extent to which a candidate cause was "the" cause of the outcome. To generate predictions from the CSM, we simulated virtual participants doing the task. In each clip, the model first watches the clip and records the 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 runs one simulation for each ball A and ball B, to determine whether balls A and B where difference-makers of the outcome. If a ball fails this test, none of the other aspects of causation are considered further. If a ball passes the test for difference-making, the model continues to test for how-causation, whether-causation, sufficient-causation, and robust-causation.

For how-causation, the CSM applies a very small perturbation to the ball's initial position and then checks whether ball E's spatial position at the end of the clip would have been different from what it was in the actual clip. If so, the candidate ball qualified as a how-cause of the outcome. Rather than yielding a continuous measure, each ball either qualifies as a how-cause or doesn't (see Table A1). Testing for how-causation captures whether there was a direct transfer of force between the candidate cause and ball E, or whether there was an indirect transfer of force, as in the causal chain where B collides with A, and A subsequently collides with E.[7]

TG: revisit footnote that explains how-causation

For whether-causation, sufficient-causation, and robust-causation, the model runs two simulations for each ball per virtual participant. This means that each virtual participant either believes that the candidate cause passed a particular test (if both simulations yield a positive outcome), failed a test (if both simulations yield a negative outcome), or is unsure (if the result of one simulation was positive, and the result of the other was negative). The model has two free parameters that affect its predictions: a NOISE parameter which affects the extent to which a ball's motion is perturbed when counterfactuals are simulated, and a PERTURB parameter which captures how much the initial location of the alternative cause is perturbed when testing for robust-causation.

To test for whether-causation, the model removes the candidate cause shortly before it would have participated in its first collision. It then simulates how the world would unfold without the cause being present. To capture people's uncertainty in what would have happened, the model applies noise to the movements of the two remaining balls. It applies this noise from the timepoint onwards at which the first collision with the candidate cause occurred. It is at this point in time that the counterfactual world deviates from the world that actually happened. 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 was removed from the scene, then ball A qualifies as a whether-cause of E's going through based on this simulation. In contrast, if ball E would have gone through the gate even if ball A had been removed from the scene, then ball A didn't qualify as a whether-cause.

TG: check the tense here

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 sufficient for 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

[7] It would be possible to construct a more general measure of how-causation that captures the extent to which the outcome in the counterfactual world (where the ball was randomly perturbed) differs from the outcome that actually happened. For example, this measure could be sensitive to the spatiotemporal distance between the actual outcome and the counterfactual outcome. We would get a continuous notion of how-causation by defining a cutoff in that space (e.g. the spatiotemporal distance that would be required to classify an outcome event as different from the one that actually happened), and then checking for each sample on what side of the cutoff the outcome event fell. Here, however, we will use this simpler notion of how-causation.

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 from the time point onwards at which the candidate cause participated in its first collision.

TG: double check here: so far, the global version of the model only applies noise to the target ball (not to the cause ball); the local version might be better here

Finally, the model considers a situation in which the candidate cause had also been removed from the scene and records the outcome in this situation. Ball A qualifies as a sufficient-cause of E's going through the gate if a) ball E would still have gone through the gate in a situation in which ball B had been removed, and b) 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 what the outcome would have been if the alternative cause had been perturbed and the candidate cause had been removed. Again, noise is applied to each ball's motion from the timepoint onwards at which the candidate cause participated in its first collision. A ball qualifies as a robust-cause if the outcome in the counterfactual situation where 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.

Using this procedure, we simulated 41 virtual pariticipants to match the number of participants in our experiment. To get the model's predictions for participants' mean judgments in the experiment, we first averaged over our virtual participants the predictions for the different aspects of causation. It turned out that the aspects of whether-causation and robust-causation were very highly correlated for the stimuli we considered in this experiment. Given that Experiments 1 and 2 have already established the importance of whether-causation, we will continue to use whether-causation here but not consider robust-causation any further for now.

TG:

- report how high the correlation actually is?
- say something about the cases for which it comes apart?
- maybe say something about discussing it in the GD

We will compare three version of the CSM. Each model first considers whether a cause was a difference-maker, and only then tests for the different aspects of causation (cf. Figure 6). The CSM_{W} only considers whether-causation, the CSM_{WH} considers whether-causation and how-causation, and the CSM_{WHS} considers whether-causation, how-causation, and sufficient-causation. 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 run a linear regression that adjusts the weights on each aspect of causation (cf. Equation 7). The CSM has one free parameter for the degree of noise that is applied to each balls' movement in the counterfactual simulations, and one parameter each 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.

TG:
- what is the correct number of parameters in our model?
- how does it compare with the heuristic model?

- parameters in the model:
  - noise
    - when does it start
    - who does it apply to
    - different noise for the different tests?
  - number of simulations
  - number of simulated participants
  - degree of perturbation

Participants. 41 participants ($M = 33.7$ years, SD = 10.5, 21 female) took part in this experiment.

Design and Procedure. The experiment featured the 32 clips shown in Figure 15. The order in which the 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". Each slider could be set independently. Hence, it was possible to give a low rating for both balls, or a high rating to both balls. On average, it took participants 21.2 (SD = 4.96) minutes to complete the experiment.

Results. Figure 17 shows participants' mean causal responsibility judgments for each of the two balls in the 32 clips. It also shows the predictions of the CSM_{WHS} which considers whether-causation, how-causation, as well as sufficient-causation.

Figure 18 shows how well each of the three versions of the CSM account for the data overall. Table 2 shows the parameters for the different models and Table A1 shows each model's prediction for the all the different clips (as well as the values of the different aspects of causation). A model that only considers whether-causation leaves much variance unaccounted for. A model which in addition to whether-causation, also takes into account how-causation explains participants' judgments significantly better, $F(1, 61) = 36.38, p < .001$. A model that also considers sufficient-causation does even better $F(1, 60) = 21.09, p < .001$. Table 2 shows how much weight each version of the counterfactual simulation model puts on the different aspects of causation. To further evaluate model fit, we also performed crossvalidation (see Table 3).

TG: update these values; or don't report and just use BIC?

DL: more detail here; what are the F-tests on; consider replacing with likelihood tests (like we did in the goalie paper); run regressions on the level of individual participants

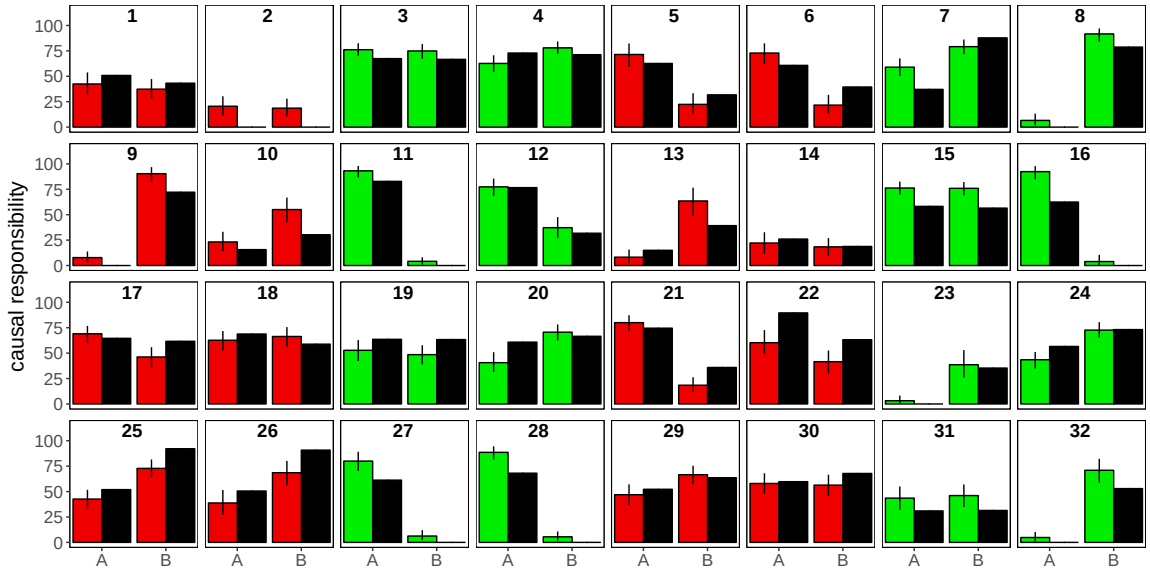


Figure 17. Mean causal responsibility (red = negative outcome, green = positive outcome) and model predictions by CSM_{WHS} (black bars). *Note*: Error bars indicate bootstrapped 95% confidence intervals.

Table 2
Regression results for the four different version of the counterfactual simulation model *(CSM). The table shows for each model, how much weight is put on the different aspects of causation. It also shows summary statistics for how well each model accounts for participants' mean judgments.* Note*: The standard error of each estimator is shown in parentheses.*

	CSM_{W}	CSM_{WH}	CSM_{WHS}	CSM_{WHSR}
whether	93.41*** (5.29)	46.19*** (6.63)	45.27*** (5.76)	28.58** (10.17)
how		36.40*** (4.28)	27.89*** (4.15)	31.68*** (4.49)
sufficient			25.64*** (5.58)	24.29*** (5.49)
robust				14.87 (7.56)
Pearson's r	.71	.83	.88	.88
Spearman's ρ	.69	.76	.84	.86
Res. Std. Error	23.41 (df = 63)	16.03 (df = 62)	13.93 (df = 61)	13.61 (df = 60)
F Statistic	311.98***	368.83***	332.82***	262.35***

*p<0.05; **p<0.01; ***p<0.001

TG: update table to remove robustness

To get a better sense for why the different aspects of causation are required in order to adequately explain participants' judgments, let us take a look at the five cases that we already discussed in the introduction above. Figure 19 shows these clips together with participants' mean judgments as well as the predictions of the different version of the CSM.

The counterfactual simulation model which only considers whether-causation as a predictor (CSM_{W}) has trouble accounting for several aspects of the data. For example, for the causal chain, it cannot capture that people give a relatively high rating to ball

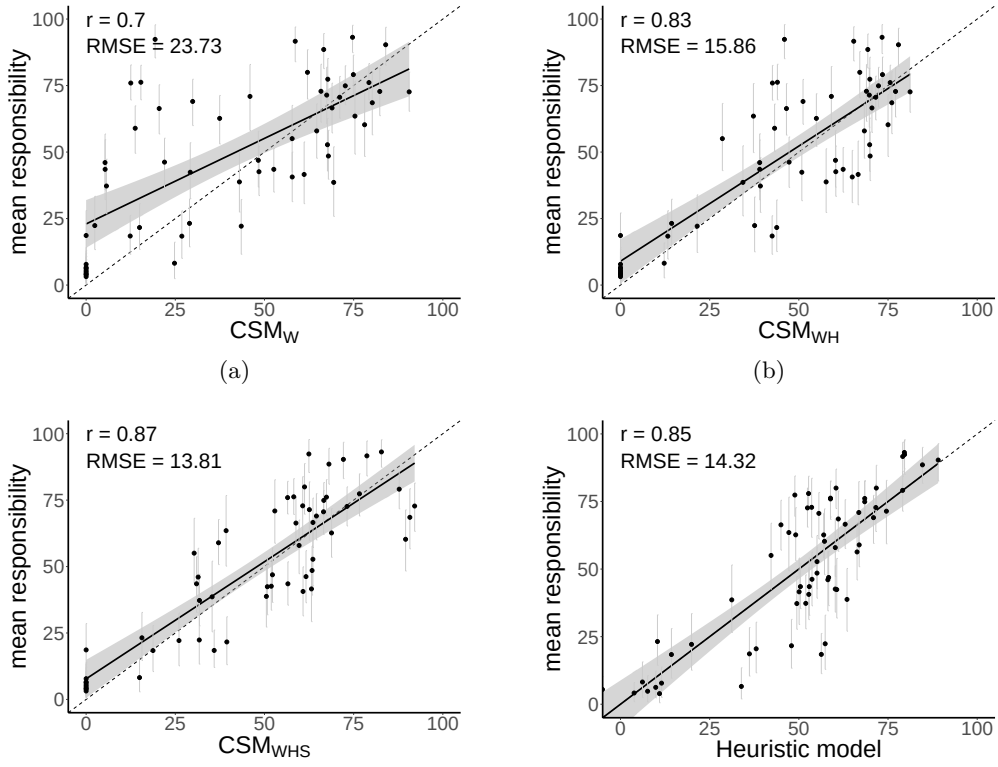


Figure 18. Scatter plots of different versions of the counterfactual simulation model (CSM), as well as the heuristic model. *Note*: W = whether-cause, H = how-cause, S = sufficient-cause. Error bars indicate 95% confidence intervals.

TG: the heuristic model predicts two values out of range (i.e. less than 0); mention this? run different regression models that assume a constrained dv, such as a logistic regression?

Table 3

Crossvalidation results. Note*: The r and root mean squared error (RMSE) columns show the median correlation and RMSE for 1000 split-half crossvalidation runs. The values in parentheses show the 5% and 95% quantiles of the distribution. The BIC scores are based on running the regression models on the full data set. For BIC scores, lower values indicate better model fit.*

model	r	RMSE	BIC
CSM_W	0.705 (0.60, 0.82)	23.49 (18.71, 27.47)	592.53
CSM_{WH}	0.819 (0.75, 0.88)	16.52 (13.81, 19.10)	547.20
CSM_{WHS}	0.864 (0.81, 0.90)	14.85 (12.71, 17.18)	532.31
Heuristic	0.798 (0.69, 0.87)	16.81 (14.09, 20.30)	559.73

DL: table needs to be referred to in text

A even though it was not a whether-cause of E's going through the gate. As Table A1 shows, $P_W(A, e)$ is only 15, whereas $P_W(B, e)$ is 85. This means that participants generally

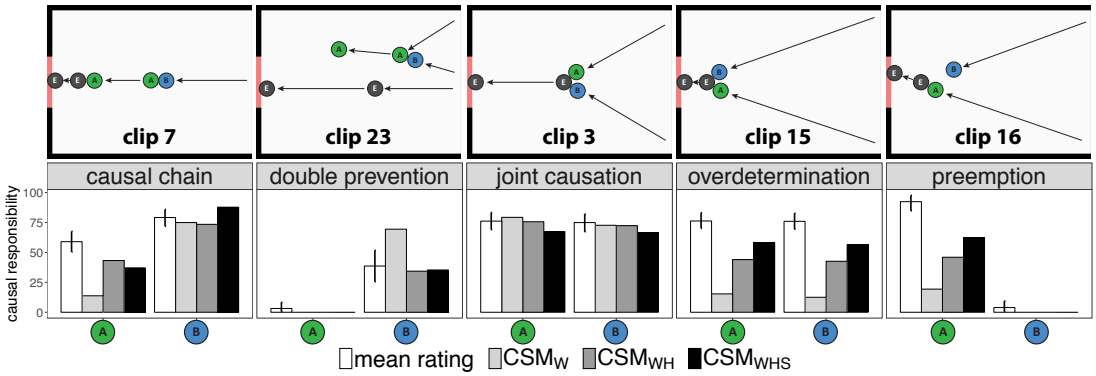


Figure 19. Participants' mean causal responsibility judgments (white bars) for the same selection of clips as shown in Table 1. The gray bars show the predictions of different version of the counterfactual simulation model. *Note:* Error bars indicate bootstrapped 95% confidence intervals.

TG:

- pick out some prevention cases as well?
- maybe make clear here that all models first use DM to select candidate causes
- increase spacing between plots and legend below

believed that ball E would still have gone through the gate even if ball A hadn't been present in the scene, but that it wouldn't have gone through the gate if ball B hadn't been present. The CSM_{W} also underpredicts participants' judgments in the overdetermination and the preemption case. Here, again, ball E would still have gone through the gate if either of the balls had been removed from the scene. The only way for this model to assign any responsibility in these cases is by assuming a high baseline rating (i.e. the model's intercept). However, this leads the model to make exaggerated predictions for ball A in the double prevention case, and ball B in the preemption case. The model also overpredicts ratings to ball B in the double prevention case. Even though participants know that E would not have gone through the gate if ball B hadn't been present in the scene, $P_W(B, e) = 74$, participants still give a fairly low rating to ball B.

The CSM_{WH} takes into account both whether-causation and how-causation to predict participants' judgments. Taking into account how-causation helps resolve many of the problems that the simple CSM_{W} faced. For example, it better captures participants' judgments for the causal chain. By taking into account that both balls were how-causes of E's going through the gate, and that B was additionally also a whether-cause, the CSM_{WH} can explain this pattern of results without assuming an overly high base rating. Furthermore, it can account for the fact that ball B's rating in the double prevention case is relatively low since it wasn't a how-cause of the outcome. Considering how-causation also allows the model to predict higher ratings for the overdetermination and the preemption case. For example, even though in the overdetermination case, neither ball A nor ball B were a whether-cause of E's going through the gate, they were both how-causes. While the CSM_{WH} does quite a bit better than the CSM_{W} overall, there are still some patterns that the model struggles with.

In particular, the CSM_{WH} predicts that participants will give a higher causal rating for the joint causation case (where both balls are whether-causes) than the overdetermination case (where neither ball is a whether-cause).

By taking sufficiency into account (in addition to whether-causation and how-causation), the CSM_{WHS} can better account for participants' judgments in the joint causation and overdetermination case. In joint causation, both balls are whether-causes but neither of the balls is sufficient for making ball E go through the gate. The opposite holds for the overdetermination case. Here, neither ball is a whether-cause but both balls are individually sufficient for making E go through the gate. By acknowledging that participants care about both aspects of causation, the model correctly predicts that the ratings should be similarly high in both cases. That the CSM_{WHS} still predicts a slightly higher rating for the joint causation case than the overdetermination case comes from the fact that overall, participants put more weight on whether-causation than sufficient-causation in their judgments (cf. Table 2).

Even though the 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. In contrast, the model's prediction is lower than people's judgments. This is the case since ball A doesn't qualify as a whether-cause in this case (unlike, for example ball B in the causal chain for which the model correctly predicts a high rating). We will discuss this case in more detail in the General Discussion.

Heuristic model. So far we have focused our analysis exclusively on the CSM. We have seen that the simple CSM_{W} does a very good job explaining participants' causal judgments in Experiments 1 and 2. Experiment 2 further showed that counterfactual contrasts are necessary for making sense of causal judgments.

The more complex CSM_{WHS} accounts well for participants' judgments in Experiment 3. However, we don't know yet whether a model that relies just on information about what actually happened might also explain participants' judgments in these more complex cases. It is possible, in principle, that participants rely on different strategies to make causal judgments in simple cases (which only feature a single candidate cause) versus complex cases (which feature more than one candidate cause).

White (2014) has suggested a number of clues that a person may use to make causal judgments about singular events. The clues to causation are derived from the hypothesis that our original source of causal knowledge stems from the experience of acting on objects (White, 2009, 2012a). From these direct experience, we abstract features of causal interactions that serve as heuristics for identifying causal relationships. Events will be seen as causal to the extent that they resemble the features derived from experiencing actions on objects.

Table 4 summarizes the different features and gives a short description for what information each feature captures. The features serve as heuristics since they do not necessarily identify causal relationship correctly, but serve as guides for identifying causal relationships under conditions of uncertainty. The heuristic predicts that the more features are true about a particular event, the more likely it is judged to be 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 they believed that

Table 4
Singular clues to causality in causal judgments as discussed in White (2014). Note*: The "variable name" column refers to the features as shown in Table 5.*

#	feature	description	variable name
1	human action	Does a human agent act on an object?	
2	two perceived objects	Do two objects interact?	
3	prior activity of actor	Do A or B move initially?	prior movement
4	initially passive patient	Does E move initially?	initial movement of E
5	direct contact between agent and patient	Do A or B contact E?	contact with E
6	monodirectional influence	Is the influenced perceived to be monodirectional?	
7	change in patient upon contact	In what way do A or B change E?	change of E's speed; change of E's movement direction
8	property transmission	Do A or B transmit force to E?	transfer of force
9	brief duration of interaction	Is the interaction between A or B and E brief?	
10	occurrence of a force impression	Do A or B create a force impression?	
11	occurrence of a causal impression	Do A or B create a causal impression?	
12	amount of perceived force exerted by the cause	How much force do A or B exert on E?	change of other objects' speed
13	outcome magnitude information	How much do A or B change E?	change of E's speed; change of E's movement direction
14	cross-modal correspondences	Was there corresponding evidence from different modalities?	
15	exclusivity	Do A or B exclusively affect E?	exclusive contact with E

the event is a causal relation or not. There was a high correlation between the number of causal features of an event, and participants' cause judgments.

There are some features that do not apply for the interactions between the billiard balls considered here, such as whether human action was involved (#1), or whether there was corresponding evidence from multiple modalities (#14). There are also some features that do not discriminate between the different clips that we showed to participants, such as whether two perceived objects interacted (#2), or whether there was brief duration of causal interaction (#9). However, there are a number of features that do differ between the clips. Table 5 shows eleven different features that we coded for our clips and briefly describes them.

Table 5
Features of the heuristic model of causal judgment. The "implementation" column explains how each variable was implemented.

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

We defined a Heuristic model that uses a linear combination of these features to predict participants' judgments. Starting with an initial model that contains all the features, we used the `stepAIC()` function of the R package `MASS` (Venables & Ripley, 2002) to find the regression which best trades off complexity and model fit. The results of this regression are shown in Table 6. Overall, the heuristic model achieves a good fit to participants' judgments (cf. Figure 18e). However, note that with 8 free parameters, the heuristic model has twice as many free parameters as the CSM_{WHS} and still performs worse. The fact that the heuristic performs relatively poorly in the crossvalidation (see Table 3) suggests that it may be overfitting the data, and that the set of specific parameter values is unlikely to generalize well across different situations.

It is also worth noting that some of the predictors have a negative sign. For example, judgments to ball A and B *decreased* when these balls were moving initially, and when E's speed increased with the collision. Furthermore, judgments to A and B *increased* when ball E was moving initially. For these features, the heuristic model as discussed in White (2014) predicts the opposite relationships.

Table 6
Results of a regression model that uses the subset of features listed in Table 5 which jointly best explain the data. Note*: The standard error of each estimator is shown in parentheses.*

intercept	7.39 (10.61)
initial movement of A/B	-49.05^{**} (15.52)
initial speed	22.74^{***} (6.46)
change of E's speed	-15.31^{***} (4.23)
change of other objects' speed	11.37^{**} (4.14)
transfer of force	44.84^{***} (5.80)
initial movement of E	19.06^{**} (5.52)
exclusive contact with E	21.14^{***} (5.96)
Pearson's r	.85
Spearman's ρ	.78
Residual Std. Error	15.31 (df = 56)
F Statistic	20.02^{***}
Note:	*p<.05; **p<.01; ***p<.001

TG: double check what these predictions look like when running separately for causation vs. prevention

Individual differences. So far, we have only looked at participants' mean judgments. For Experiments 1 and 2, the CSM has only one way of predicting individual differences: participants should come to different causal judgments depending on what they believe would have happened if the candidate cause hadn't been present in the scene. For Experiment 3, the CSM has multiple ways of capturing individual differences. Participants may differ in their belief about whether the candidate cause was a whether-cause, a sufficient-cause, and how robust it was.[8] Additionally, the CSM can accommodate individual differences by assuming that participants may differ in how much they take the different aspects into account when making causal judgments.

To analyze the extent to which participants differed in how their causal responsibility judgments, we ran our model on each individual participant's responses. For each participant, we looked at how well their judgments were explained by the four different versions of the CSM: CSM_{W}, CSM_{WH}, CSM_{WHS}, and CSM_{WHSR}. We also included a baseline model which simply uses a participant's average response to predict their judgments. We assigned participants to the different models based on BIC scores which take into account both model fit and model complexity.[9] Table 7 shows the results. Almost none of the participants' responses are best explained by a model that only considers whether-causation. The

[8]Participants may also differ in their belief about whether the candidate was a how-cause even though this aspect of causation is easier to assess than the other aspects of causation. Remember that for how-causation, we simply need to keep track of collisions between balls.

[9]Models were excluded that had a negative sign on any of the predictors.

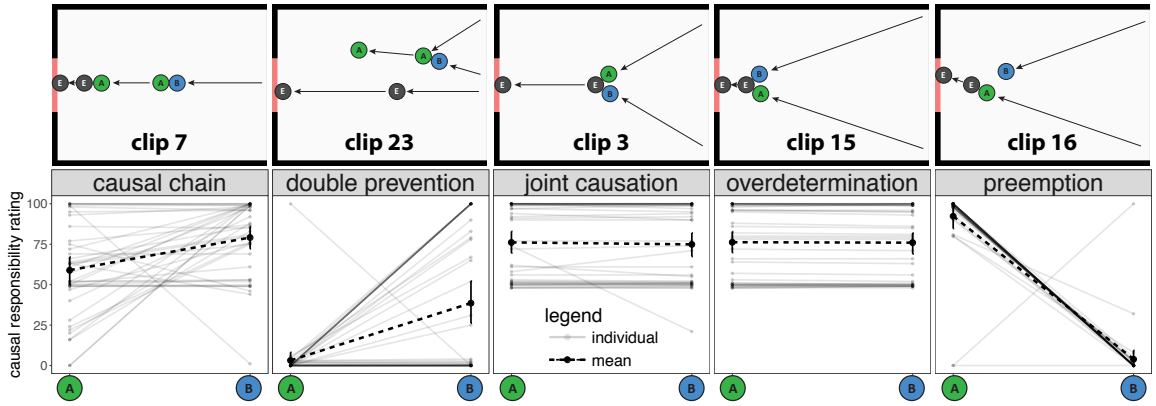


Figure 20. Plots illustrating individual differences for a selection of the clips shown in Experiment 3. *Note*: Thin gray solid lines indicate individual participants' ratings for balls A and B. The thick dashed line shows group averages. The error bars indicate bootstrapped 95% confidence intervals.

TG: set of situations considered here ok? or shall i add a case in which judgments are flipped (i.e. for one group $A > B$ and for the other group $A < B$?

TG: merge with the figure on model predictions?

Table 7
Number of participants whose causal responsibility judgments were best explained by the different models. Note: *The baseline model uses participants' mean rating to predict their response in each case.*

	baseline	CSM_W	CSM_{WH}	CSM_{WHS}	CSM_{WHSR}
# participants	2	2	13	18	6

majority of participants' responses are best captured by the CSM_{WHS} and the CSM_{WH}. Six participants' judgments were best explained by a model that includes robust-causation as an additional factor.

Figure 20 shows individual participant judgments for the selection of five clips which we already discussed above. Each gray line in the plot shows what judgments a participant gave for ball A and ball B. The dashed line in each plot indicates the averaged response. For many situations, there was considerable variance in participants' judgments.

For example, in the "causal chain" (clip 7), many of the pariticipants gave a higher rating to ball B than ball A. However, there was also a group of participants who judged both balls equally, either giving a rating of around 50 or 100 to each. There was also one participant who gave a much greater rating to ball A than ball B. We can make sense of this interindividual variation by assuming that participants differ in how much how-causation and whether-causation affects their judgment. Participants who mostly care about how-causation will gave identical ratings to both balls. However, participants who care more about whether-causation (and sufficiency) will gave a greater rating to ball B than ball A.

In case of the "double prevention" (clip 23), there were again two broad groups of participants. Both groups saw ball A as not having been responsible at all for E's going through the gate. However, they differed in how much causal responsibility they assigned to ball B. Whereas one group gave a close-to-zero rating for ball B as well, another group gave higher ratings to ball B. 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 again have differed in how they weigh how-causation and whether-causation. Participants for whom how-causation is critical will have given a zero rating to ball B. Participants who gave higher ratings to ball B may have done so because they care about whether-causation to some extent. Second, some participants may have simply been uncertain as to whether ball B was actually a whether-cause of E's going through the gate. The counterfactual judgments for this clip show that participants weren't completely sure that ball E would have missed the gate if ball B hadn't been present in the scene (cf. Figure 16).

For both the "joint causation" case (clip 3) and the "overdetermination" case (clip 15), there were two distinct groups of participants. One group gave a 50 rating, and the other a 100 rating to each ball. As discussed above, the CSM accounts for this pattern of results by assuming that participants differ in how much they care about whether-causation and sufficient-causation. Participants who cared more about whether-causation gave a high rating for the "joint causation" case and a lower rating for the "overdetermination" case. Participants who cared more about sufficient-causation show the reverse pattern.

TG: add any stats here to support this point?

The clips discussed sofar suggest that there was considerable variance in how causal judgments were reached. The results for the "premption" clip show that this need not be the case. Indeed, the pattern of results here was strikingly consistent: almost all participants gave a close-to-maximal rating to ball A and a zero rating for ball B. As discussed above, the CSM struggles with explaining this pattern since A was not a whether-cause of E's going through in this case, but whether-causation is an important predictor or how people make causal judgments across a variety of situations. One possibility is that people consider the different aspects of causation to differing degress based on the situation. If that's the case, we need a theory that explains to what extent the different aspects matter in different situations. We will return to the problem of preemption in the General Discussion. Overall, we saw that there was both variance between participants in how much they cared about the different aspects of causation, as well as in how much the different factors affected participants' judgments in different situations.

DL: How would we deal with the comment that perhaps people invoke different aspects according to the nature of the clip itself - so trying to fit a single model for every case obscures what's really going on

TG: I'm mentioning this possibility here now; not sure at this point how to really tackle the problem

Discussion

Experiment 3 provided a challenging test ground for the counterfactual simulation model. By looking into situation with two candidate causes, we were able to reconstruct

many of the situations that have troubled counterfactual theories of causation, such as overdetermination, and preemption. The results show that CSM was up to the challenge. The model was able to account for participants' causal responsibility judgments to a high degree of quantitative accuracy. Whereas in Experiments 1 and 2, we only needed to consider whether-causation in order to explain participants' judgments, Experiment 3 highlights that participants are sensitive to different aspects of causation when making causal judgments. First, we need to consider whether or not a candidate qualifies as *a* cause of the outcome. To do so, the CSM considers whether anything about the outcome of interest would have been different in a situation in which the candidate cause had been removed from the scene. For each candidate that has qualifed as *a* cause of the outcome, the CSM then determines to what extent the candidate was *the* cause of the outcome by considering the following four aspects of causation:

1. **whether-causation**: Did the presence of the candidate cause made a difference to whether ball E went through the gate?
2. **how-causation**: Would the outcome event (finely construed) have been different if the candidate cause had been somewhat perturbed?
3. **sufficient-causation**: Would the candidate cause still have brought about the outcome in a situation in which the alternative causes had been absent?
4. **robust-causation**: Would the candidate cause still have brought about the outcome in a situation in which the alternative causes had been somewhat perturbed?

Overall, a model that considered whether-causation, how-causation, and sufficient-causation struck the best balance between model complexity and fit. However, there were also a number of participants whose answers were best explained by assuming they took all four factors into account.

TG: not sure about the formulation of sufficient-causation and robust-causation here

We compared our counterfactual simulation model with a 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 explain how people make causal judgments (cf. White, 2014). Even though the best-fitting version of this model contained twice as many free parameters as CSM_{WHS}, it still performed worse. Overall, this shows that a model which tries to do away with counterfactual contrasts fails to adequately explain participants' causal responsibility judgments. In Experiment 2, we compared situations that were identical in terms of what actually happened and only different in what would have happened if the candidate cause hadn't been present. The results of this experiment showed that people's causal judgments cannot be captured without counterfactuals. However, 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. But again, the results of Experiment 3 show that our counterfactual simulation model better explains participants' judgments than a model that only looks at what actually happened.

We also looked at individual participants' judgments and saw that there was considerable variation. The CSM explains this variation as systematic differences in how much

participants take the different aspects of causation into account when making causal judgments. Currently, we can only speculate where these individual differences come from. One possibility is that participants merely 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 the gate instead of *not* going through the gate. Another, more intriguing, 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 the "how" versus "whether" to show up in other domains as well. For example, people for whom a more direct connection is critical for causation (as revealed by test for how-causation), might differentiate more strongly between acts of omission versus commission (McGrath, 2005; Spranca, Minsk, & Baron, 1991; Stephan, Willemsen, & Gerstenberg, 2017), or pay particular attention to the role of force in harmful events (Cushman & Young, 2011; Greene et al., 2009; Iliev, Sachdeva, & Medin, 2012; Mikhail, 2007).

TG: this section might still need some work

General Discussion

TG: add a section that talks about the eye-tracking paper; computational vs. process level explanation

This paper introduced a novel model of how people make causal judgments: the counterfactual simulation model (CSM). The CSM is the first model to accurately predict causal judgments about dynamic events in physical scenes. A key claim of the model is that causal judgments are intimately linked to counterfactuals (cf. Gerstenberg, Peterson, et al., 2017). This claim is not new (cf. Hart & Honoré, 1959/1985; Lewis, 1973; Lipe, 1991; Mackie, 1974). However, previous models that tried to link causal judgments to counterfactuals have been 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 a rich counterfactual model that considers multiple contrasts, handles situations that have troubled previous accounts, and that it does so in a way that bridges previous conceptions of causation.

The CSM postulates that people's causal judgments about particular events are based on a generative model of the scene which supports the simulation of what would have happened in different counterfactual situations. This generative model dictates the causal processes that govern how the world unfolds. For the dynamic collision events considered here, people's understanding of the situation can be expressed as an intuitive model of physics. Based on previous work, we assume that people's intuitive understanding of physics is in important ways similar to a game engine that can be used to generate physically realistic stimuli (cf. 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 particular events hadn't happened, if particular objects hadn't been present in the scene, or if something about these objects had been changed (see also Chater & Oaksford, 2013; Gerstenberg & Tenenbaum, 2017; Goodman, Tenenbaum, & Gerstenberg, 2015).

Traditionally, formal models that implement counterfactual theories of causation have focused on relatively abstract models based on structural equations that express the relationships between different variables, whereby each variable indicates whether or not a particular event occurred (Halpern, 2016; Halpern & Pearl, 2005). In this paper, we have proposed a different strategy. Instead of representing people's causal model of the situation in terms of 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). Different aspects of causation can then be probed by considering different kinds of counterfactual contrasts over this representation. We get a notion of WHETHER-CAUSATION by simulating whether the outcome would have been qualitatively different if the candidate cause hadn't been present in the scene. Most counterfactual theories have focused on this aspect of causation. However, by casting people's causal representation of the scene in terms of a physics engine that supports counterfactual intervention, we can 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 also get a notion of SUFFICIENT-CAUSATION and ROBUST-CAUSATION. A cause is sufficient if it would still have brought about the outcome even if all alternative causes had been removed from the scene, and robust to the extent that it would have brought the outcome even if the alternative causes had been somewhat perturbed. Overall, the CSM combines two different kinds of counterfactual interventions, "remove" and "change", with two different ways of specifying the granularity of the outcome, "coarse" (the outcome of interest happened or didn't happen) and "fine" (the outcome happened at a particular point in time and space), to yield different aspects of causation that affect people's causal judgments.

TG: mention as a factor what the targets of the intervention are?

The results of three experiments support the CSM. In Experiment 1, the model captures participants' causal judgments about simple collision events to a high degree of quantitative accuracy. Participants' cause and prevention judgments increase the more certain they are that the outcome would have been different if the candidate cause had been absent. Experiment 1 featured a broad range of clips whereby what actually happened was different in each clip. Thus, it could in principle be possible to explain people's causal judgments just in terms of what actually happened and without reference to counterfactual contrasts (although see Gerstenberg, Peterson, et al., 2017, for eye-tracking evidence that participants spontaneously engauge in counterfactual simulation in these clips).

In Experiment 2 we created pairs of clips in which what actually happened was identical, but what would have happened in the relevant counterfactual 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. As predicted by the CSM, participants' judgments differed

Table 8
Qualitiative 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	no	yes	no	yes
considers processes	yes	no	yes	yes
handles multiple causes	yes	yes	no	yes
generalizes beyond physical causation	somewhat	yes	no	somewhat

markedly between pairs of clips as a function of what would have happened, even though what actually happened was held fixed.

Finally, in Experiment 3, we looked at situations that featured two candidate causes. In this more complex setting, we were able to reconstruct many of the situations 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 weight they give to different aspects when making causal judgments. A heuristic model based on features that only capture what actually happened (cf. White, 2014) didn't account for participants' judgments as well.

TG: briefly compare the different models to each other in text

In the remainder of this paper, we will consider open challenges and future directions for the CSM.

TG: maybe have a brief summary of what's to come here

Future directions and open challenges

Causal relata: Objects vs. events. Most philosophical theories of causation take the causal relata to be events – cause events bring about effect events (cf. Paul & Hall, 2013). For example, ball A's colliding with ball B is what caused ball B's going through the gate. However, in natural language, it is often more common to talk about objects (or agents) as having caused particular events. We say "Tom broke the vase" rather than "Tom's hitting the vase caused the vase to break", or "a rock smashed the window" rather than "the collision between the rock and the window caused the window to smash" (cf. Talmy, 1988).

TG: add some linguistics references here

The structural equation account of causal judgment discussed in the introduction (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 defines counterfactual contrast as 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 counterfactual situations. While there are many ways to bring about a counterfactual world 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, ...), the world in which ball A hadn't been present is well-specified.

More generally, we believe that game engines provide a natural starting point for exploring what kinds of counterfactuals people may bring to mind when imagining how things could have turned out differently (Ullman et al., 2017). In a game engine, like the one that we used to generate our stimuli, some interventions 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. In contrast, we cannot directly intervene on events such as the collision between A and B.

TG: incorporate the point about intervening on attributes vs. relations (as discussed with josh)

Note that we share this focus on objects as the targets of analysis with Wolff's (2007) force dynamics model discussed in the introduction. The force dynamics model considers the forces associated with the agent and patient to determine the extent to which the agent brought about a certain outcome involving the patient. 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. According to White (2009), the asymmetric way in which we experience ourselves as agents (exerting forces) rather than patients (resisting forces) forms a template that subsequently biases our perception of forces, such that in a simple Michottean interaction of two balls, we see the moving ball as exerting more force on resting ball (and launching it) than vice versa (White, 2006, 2017, but see also Mayrhofer & Waldmann, 2016; White, 2011 for possible dissociations between force and cause judgments).

By defining interventions on objects rather than events, we can also make sense of other physical concepts that are related to causation. In recent work (Gerstenberg, Zhou, Smith, & Tenenbaum, 2017), we have started to explore people's intuitive understanding of physical support (cf. Battaglia et al., 2013; Hamrick et al., 2016). In our experiments, participants viewed 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 of interest were removed. The results showed that there was a very close correspondence between the responsibility judgments of one group and the predictions of another group about what would happen if the block was removed. The greater the proportion of blocks that was predicted to fall off the table, the more responsible that block was seen. Alternative models that tried to explain 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 considering physical support, the scene is static and there are no events.[10] Since there are no events a model of physical support cannot operate on events. Instead, by considering interventions on objects rather than events, we can handle both judgments about dynamic causation and static support in a unified way.

TG: add references to philo literature – ask Laurie

The language of causation. We use many different causal verbs to describe what happened (Abelson & Kanouse, 1966; Brown & Fish, 1983; Rudolph & Forsterling, 1997). For example, we distinguish between "causing", "enabling" and "helping" Lombard (1990);

[10] At least no events in the intuitive sense describing changes of states. In philosophy, events are often construed more broadly such that "a block lying on the table" would count as an event.

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 a causal Bayes net representation (Sloman, Barbey, & Hotaling, 2009). Wolff's (2007) force dynamics model differentiates not only "caused" from "enabled" but also models other causal verbs such as "prevented" and "despite" based on the force configurations that these verbs map onto (cf. Figure **??**). In the force dynamics model, "enabled" is distinguished from "caused" by way of the patient's tendency and the alignment between agent force and patient force. While in the case of "caused", the patient's force didn't point toward the endstate, and the agent and patient forces didn't point in the same direction, in the case of "enabled" the patient's force *was* directed toward the endstate, and the agent and patient forces pointed in the same direction.

The CSM suggests a new perspective on the semantics of different causal verbs (cf. Gerstenberg & Tenenbaum, 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 verbs 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 "helped". Accordingly, "caused" applies best when both aspects of causation are high. "Enabled", in contrast, is specifically sensitive to whether-causation but does not require how-causation. If A moves B out of the way such that E can go through the gate, A enabled (but didn't cause) E to go through the gate (cf. Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017; Sloman et al., 2009). "Helped" is the weakest out of the three terms. In order to have helped, it suffices to have been a how-cause of the outcome. For example, if E is already headed toward the gate and A collides with E to speed it up, then A helped (but didn't cause or enable) E to go through the gate. While the force dynamics model doesn't draw a distinction between "enabled" and "helped" (both verbs map onto the same force configuration), our account suggests a way of differentiating between the terms. In preliminary experiments, we have found support for the role that whether-causation and how-causation play for people's descriptions of causal interactions (see also Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012). Because the CSM defines causal concepts in terms of counterfactual contrasts, it also naturally captures intrinsically counterfactual statements such as "almost caused" or "almost prevented" (Gerstenberg & Tenenbaum, 2016).

Causation by omission. Omissions have a difficult causal status. Consider the classic example: My gardener's not watering my plants caused them to die. When considering omissions as causes we run into two problems: (1) the problem of causal selection, and (2) the problem of underspecification. The problem of causal selection has received most attention to date (Bello & Khemlani, 2013; Bernstein, 2014; Halpern & Hitchcock, 2015; Hesslow, 1988; Hitchcock & Knobe, 2009; McGrath, 2005; Tang, 2015; Willemsen, 2016; Wolff et al., 2010). Why was it the gardener and not the Queen of England that caused my flowers to die? If the Queen of England had watered the plants, they would not have died either. While this counterfactual statement is true, it strikes us as irrelevant (Kominsky et al., 2015). One popular strategy to justify why the gardener and not the Queen of England caused the plants to die, is to rely on the role of norms and expectations. While it was the gardener's duty to water the plants, and he was expected to do so, the Queen of England was not responsible for the plants. Several empirical studies have confirmed the

role that norms play for judgments about omissions (Clarke, Shepherd, Stigall, Waller, & Zarpentine, 2015; Henne, Pinillos, & De Brigard, 2016; Livengood & Machery, 2007). For example, when two cars collide at an intersection whereby one driver had a red light and the other one a green light, participants judge the *not stepping on the brakes* of the driver with the red light to be more causal (Clarke et al., 2015).

The problem of underspecification, however, has received less attention. Assuming that the gardener (or someone else) had actually watered the plants, would the plants have survived? Plants die if they receive too little or *too much* water. While the problem of underspecification may not appear too pressing in the case of watering plants, it becomes more apparent in other situations. Generally, it is more straightforward to say what the relevant counterfactual is for an event that actually happened – namely, that particular event not happening. However, when nothing happened, it is unclear what the relevant counterfactual event should be. If Joe shot Steve, it is easy to imagine what would have happened if Joe hadn't shot Steve. In contrast, if Joe didn't shoot Steve, it is less clear what would have happened if Joe had shot Steve. Where would the shot have landed? Would Steve have died, or would he have survived despite being shot? Counterfactual theories require a contrast between what actually happened and what would have happened in the relevant counterfactual situation (Schaffer, 2005). While for things that happened, this contrast tends to be well-defined, the contrast is underspecified for omissions. This does not mean, however, that counterfactual theories cannot deal with situations of omission. In fact, the CSM provides an intuitive solution to the problem of underspecification (Stephan et al., 2017).

Consider a situation in which ball B goes through the gate while ball A just remained lying still in a corner. Did ball B go through the gate because ball A didn't hit it? The CSM answers this question by first simulating counterfactual situations in which ball A had actually collided with ball B and then recording whether B would have missed the gate, or still gone through in this case. To more certain it is that B would have missed the gate if A had hit it, the more the model agrees that the B went through the gate because A didn't hit it.

In contrast, consider now a situation in which ball A is again lying still but ball B misses the gate this time. To answer the question of whether ball B missed because ball A didn't it hit, the model considers the counterfactual in which ball A had hit ball B and checks whether B would have gone through the gate in this case. Assuming that the gate is relatively small, it is easier to make a ball that is headed toward the gate miss, than it is to make a ball that is not headed toward the gate go through. Thus, the CSM predicts that people should be more willing to agree with the statement that "Ball B went though the gate because ball A didn't hit it", than with the statement that "Ball B missed the gate because ball A didn't hit it". Indeed, this is what we found (Stephan et al., 2017). By looking at situations in which the problem of underspecification didn't arise, we further ruled out a mere preference for causation by absence over prevention by absence.

Normative expectations. Norms play a critical role for dealing with causation by omission. However, the current version of the CSM does neither incorporate norms, nor include any notion of optimality. The only way in which probabilities enter the model is in terms of the 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 knows.

Much research has shown that people's causal judgments are not only affected by their subjective degree of belief in what would have happened in relevant counterfactual situations, but also by their expectations about what will happen in the actual situation. 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 try to quantify how normality considerations affect causal judgments (Halpern & Hitchcock, 2015; Icard et al., 2017; Kominsky et al., 2015).

To explain participants' causal judgments in the experiments discussed in this paper, we did not need to incorporate normative expectations about what will happen. One reason for this might be that we generally had balanced designs. In all of the 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 considered here. One of the major drawbacks of vignette studies is that much about what actually happened is left implicit, and different participants may fill in the gaps differently. The visual paradigm allows for stringent tests of normality effects because there is little to no uncertainty about what actually happened in a particular situation.

The problem of preemption.

TG: mark asymmetries between preemptive causation and prevetion

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, ball A collides with ball E and makes it go through the gate before ball B would have done the same a moment later (cf. Figure 19). 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 to predict this high rating is that the preempting cause was not a whether-cause of the outcome, and whether-causation generally is an important factor for people's causal judgments.

Indeed, the fact that the presence of the preempted cause makes no difference to how the preempting cause is evaluated is often taken as a key piece of evidence in favor of process theories over counterfactual theories of causation (Paul, 1998,?; Paul & Hall, 2013). However, the results of the experiments reported in this paper make it clear that counterfactual contrasts are a necessary tool for explaining causal judgments. So, what can we do about preemption?

One possible solution is to argue that how important the different aspects of causation are varies between situations. However, this only pushes the problem up a level in that we

would now need a meta-level theory that predicts when certain aspects are more important than others. Alternatively, one could try to argue that people's causal representation of the situation is different from what actually happened. Remember that the CSM first identifies which candidate qualifies as "a" cause of the outcome, before it proceeds to determine the extent to which the different causes qualify as "the" cause of the outcome. At first sight, the following solution to the preemption problem 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. Since the preempted cause doesn't qualify as a cause of the outcome, 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 does not work. In Experiment 2 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 1 nor clip 5 does the brick make a difference to what actually happens (cf. Figure 12). However, the causal ratings differ dramatically between the two causes. Participants say that ball A prevented ball B from going through the gate in clip 5, but not in clip 1. 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. It's also not the case that there is simply an asymmetry between causation and prevention, in which we can remove objects that didn't make any difference when evaluating causation, but cannot do so in the case of prevention. Remember that ball A is judged to have caused B to go through the gate in clip 14 but not in clip 18. Again, the position of the brick makes a big difference to participants' causal judgments here.

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 (rather than a preventer): 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 in which there are several potential causes but only one of which actually directly affected the outcome, the other potential causes that didn't affect the outcome are removed from the causal representation of the scene. When asked to make a causal judgment about what happened, it is as if these other potential causes hadn't even been there in the first place.

Dealing with preemption is tricky. While the CSM provides some tools for tackling these cases, it doesn't yet have all the answers. More work is required to fully understand people's intuitions in situations of preemptive causation and preemptive prevention (Collins, 2000; McDermott, 1995).

TG: add another sentence about how to handle preemptive prevention?

Causal judgments vs. causal perception.

TG: CONTINUE HERE

Research in causal perception investigates what makes events look causal (Blakemore et al., 2001; Michotte, 1946/1963; Rips, 2011; Saxe & Carey, 2006; Schlottmann, 2000; Scholl & Tremoulet, 2000; White, 2012b). It often feels like we can perceive causation directly. When we see two billiard balls collide, we see that one caused the other to move (and, even

though less so, that the other caused the one to stop, Mayrhofer & Waldmann, 2014). The inference that causation happened is immediate and doesn't seem to require an explicit consideration of what would have happened if the two balls hadn't collided. So isn't the fact that we can sometimes perceive causality evidence against a counterfactual theory of causal judgment?

There are at least two ways of responding. One possibility is that causal perception and causal judgments are supported by different cognitive systems serving different cognitive functions. The function of the *causal perception system* is to provide accurate predictions for what will happen in the near future. It relies on relatively low-level features, such as spatial and temporal association that trigger a perception of causality in a bottom-up fashion. The function of the *causal judgment system* is to provide explanations for why something happened. It operates on our world knowledge that influences causal judgments in a top-down fashion – for example, through the generation of relevant counterfactual simulations.

In support of such a two-systems view, dissociations between causal perceptions and causal judgments have been reported (Levillain & Bonatti, 2011; Rips, 2011; Schlottmann, 1999; Schlottmann & Shanks, 1992; Thorstad & Wolff, 2016; Wolff & Shepard, 2013). For example, Schlottmann and Shanks (1992) demonstrated how judgments of perceived causality and whether a collision was necessary, can come apart. They showed participants variants of the Michotte launching stimuli in which they manipulated *temporal contiguity* (i.e. the time at which the second object starts moving after the collision with the first object) as well as *statistical contingency*. In the contingent condition, participants learned that the second object only moved if there was a collision *and* the second object also changed its color. If the object only changed its color but there was no collision, it didn't move. In the non-contingent condition, the second object moved as long as it changed its color (irrespective of whether there was a collision). Hence, in the contingent condition, the collision was necessary for the object to move, while in the non-contingent condition it wasn't. The results showed that participants' perceived causality ratings were strongly affected by temporal contiguity but not affected at all by whether the collision was necessary for the second object to move. In contrast, judgments of whether the collision was necessary for the second object to move were strongly affected by the contingency manipulation, whereas temporal contiguity had little effect.

A second way of responding to the challenge that causal perception poses, is to say that that causal perception and judgment are intimately related and that causal perceptions can be thought of as a particular kind of causal judgment. If this was the case, one would expect for there to be strong top-down influences on causal perception.

Indeed, Bechlivanidis and Lagnado (2013) have shown that people's causal beliefs affect their perception of the temporal order of events. Participants who had learned how a causal system worked (a puzzle solving game with several objects and non-obvious causal relationships between them), reported that they had perceived a sequence of events in an order that was consistent with how the causal system worked, but inconsistent with what they had actually seen. In another series of experiments, Bechlivanidis and Lagnado (2016) showed that the same effects occur even for very simple causal chains of three objects colliding with each other. The control condition simply shows a causal chain in which object A approaches two objects B and C which are at rest and in some distance to each

other. A collides with B, and B collides with C. Participants were asked to report the order in which the following events happened: "A started moving", "B started moving", and "C started moving". In a critical test case, participants saw a clip in which C starts moving at the same time at which A collides with B. After some temporal delay, B starts moving and stops at the position at which it would have collided with C. So in this test clip, C starts moving before B started moving. The results showed that participants mis-perceive the order of events in this clip. They tend to report the A-B-C order (which is consistent with their causal beliefs about how collisions work), instead of the A-C-B order (which is consistent with what they actually saw). So while causal perceptions and causal judgments can sometimes come apart, they still seem to be tightly linked.

What does the CSM have to say about causal perception? The CSM assumes that causal judgments are made by comparing what actually happened with what would have happened in different counterfactual situations. In causal perception experiments, researchers manipulate aspects of the causal event of interest. For example, they introduce a delay between the time at which the two objects collide, and the time at which the second object starts to move. Or they introduce a spatial gap between the two balls. Both of these manipulations affect the extent to which what actually happened is in line with what one would have expected to happen. Generally, participants tend to perceive collisions as less causal that involved a temporal delay, or a spatial gap. Participants are also sensitive to aspects of how the second ball moves, including its speed (Sanborn et al., 2013) and direction of motion (White, 2012b).

The CSM postulates that when making causal judgments, people first consider whether there was a causal connection between the candidate cause and the outcome event. To do so, we need to simulate whether the outcome event would have been any different if the candidate cause had been absent (cf. Figure 6). Note that for the cases of causal perception that we've discussed, the relevant counterfactual is trivial to compute. Indeed, it doesn't even need to be simulated – it just needs to be remembered. The object of interest would have simply stayed put. Thus, a single sample is enough to answer the counterfactual question that is needed to determine whether there was a causal connection Sanborn and Chater (2016); Vul, Goodman, Griffiths, and Tenenbaum (2014).

According to the CSM, the graded effects that certain manipulations , such as temporal delays or spatial distances, have on participants' causal perceptions arise to the extent that they cast doubt on whether there was in fact a causal connection. If the observer believes that the effect event would have been exactly the same no matter whether or not the candidate cause was present, then causal judgments should go down.

The function of causal judgments. This paper presents a descriptive account of how people make causal judgments. But what are causal judgments actually good for? It is easy to justify why we should have an accurate model of how the world works. Such a model serves as a guide for action, and helps us realize our goals efficiently. To reach our goals, we need to consider what the likely consequences of different actions on the world would be, and then plan and choose our actions in a way that is expected to work best. This process requires a generative model of how the world works, and a way of simulating the consequences of hypothetical interventions. A good causal theory is one that helps us to adequately characterize how people learn, reason, plan, and act upon the world (Woodward, 2014). But where in this process are causal judgments? Causal judgments are about events

that have already happened. How can this be useful for the future?

Causal judgments form the foundation for assigning responsibility as well as legal liability (Hart & Honoré, 1959/1985; Lagnado & Gerstenberg, 2017; Moore, 2009; Stapleton, 2008). We have to establish first that an agent's action played a causal role in bringing about the outcome before we can hold the agent responsible for what happened (Alicke, 2000; Alicke, Mandel, Hilton, Gerstenberg, & Lagnado, 2015; Allen, Jara-Ettinger, Gerstenberg, Kleiman-Weiner, & Tenenbaum, 2015; Chockler & Halpern, 2004; Gerstenberg et al., 2015; Gerstenberg & Lagnado, 2010, 2012; Halpern, 2016; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Lagnado et al., 2013; Malle, Guglielmo, & Monroe, 2014; Niemi, Hartshorne, Gerstenberg, & Young, 2016; Shaver, 1985; Weiner, 1995; Zultan et al., 2012). Holding others responsible forms a key part for regulating relationships both on the level of individuals and collectives more generally (Forsyth & Kelley, 1994; Lewis, 1948; Rai & Fiske, 2011).

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

Recently, it has been proposed that causal judgments further play the role of highlighting those aspects which not only made a difference to the outcome, but that are also likely to continue to make a difference in other situations as well (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; Meder et al., 2010). In line with the tight relationship between counterfactuals and causal judgments that the CSM postulates, it has also been found that people focus on controllable events when considering how things could have gone differently (Girotto, Legrenzi, & Rizzo, 1991).

Finally, forming causal representations of a situation also allows us to communicate efficiently what happened. For example, hearing that "Tom broke the vase", "Tom caused the vase to break", or "Tom allowed the vase to break" result in different beliefs about what happened (e.g. Freitas et al., 2017; Solstad & Bott, 2017). The CSM provides a new addition to the family of theories that tries to better understand what words people use to pick out causal events in the world (Goldvarg & Johnson-Laird, 2001; Wolff, 2007).

Going beyond physics. In this paper, we have focused on applying the CSM to making judgments about physical events. However, causality doesn't only happen between billiard balls. It happens between people, markets, countries, etc. Indeed one of the central motivations for our counterfactual account of causation was that it is flexible. Ideally, we want a theory of causation that not only applies to billiard balls but to people, too. We believe that the CSM carries the potential to be such a theory. The examples of everyday situations that we used at the very beginning of this paper comprised situations of physical, psychological, and economic causation. The CSM can be applied to a situation as long as we can represent our domain knowledge as a generative model that supports reasoning about counterfactuals. While we focused on modeling judgments about physical causation

we believe that the different aspects of causation that the CSM postulates, will help make sense of causal judgments in other domains, too, such as when social agents interact with one another (cf. Gerstenberg & Tenenbaum, 2017).

Characterizing the causal interactions between social agents requires us to develop computational models of our intuitive theory of psychology (Evans, Stuhlmüller, Salvatier, & Filan, 2017). We need such models to be able to simulate whether an agent would have acted differently if something about the person, or the situation, had been different (Fischer et al., 2016). For example, it has been shown that people can infer an agent's beliefs and desires from the actions they took Baker, Jara-Ettinger, Saxe, and Tenenbaum (2017); Baker et al. (2009). This inference can be formalized as inverse planning – people assume that an agent forms a rational plan for action to efficiently achieve its goals given its beliefs about the environment. Upon seeing an agent act, an observer can invert the agent's rational planning process to infer the underlying beliefs and desires that explain the agent's actions. An agent's desires may be simple, such as getting from A to B, or more complex, such as helping or hindering another agent (Ullman et al., 2009). Moreover, our commonsense psychology includes considerations of costs and rewards – a naïve utility calculus that incorporates not only beliefs and desires, but also preferences, and character traits (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016). For example, when two agents refused to help, children evaluate the agent more negatively for whom helping would have been easier (Jara-Ettinger, Tenenbaum, & Schulz, 2015). Recent work has also begun to explore how inferences about emotions can be expressed in computational terms Ong, Zaki, and Goodman (2015); Saxe and Houlihan (2017).

While thinking about agents is more difficult than thinking about objects, to the extent that we are able to simulate what would have happened in relevant counterfactual situations, the CSM can be used to make predictions about causal judgments. Counterfactuals not only help determine whether an action caused an outcome, they also allow us to say whether a thought caused an action. A key factor for how we evaluate others' behavior is whether they acted intentionally (Heider, 1958; Malle, 2004; Malle et al., 2014). For example, while accidental harms are indicative of clumsiness, intended harms are indicative of meanness (Uhlmann, Pizarro, & Diermeier, 2015). What distinguishes being pushed accidentally on a busy street from being pushed intentionally, is that the accidental push could have easily been avoided by stepping out of the way. However, if the push was intended, then it would still have happened even if we had stepped out of the way. Intentions make the causal relationship between actions and outcomes robust (Heider, 1958; Lombrozo, 2010).

What further complicates evaluating others is that actions often have several effects, only some of which may actually have been intended. When helping one friend move, I might have to turn down another one's request for help. Kleiman-Weiner et al. (2015) have developed a model of intention that clarifies what it means for an agent to have intended a particular outcome. In their model, intended outcomes are defined as those that make a difference to the agent's decision. They use a counterfactual criterion to define what it means for an anticipated outcome to have made a difference: if an outcome was intended this means that the agent would have acted differently in case that outcome had been different. In a number of studies, they show that people's moral evaluations of another agent's actions were influenced by what the action revealed about the agent's intentions, and by what the actual consequences were.

The CSM postulates different ways of assessing what difference a candidate cause made to the outcome. When we think about people as causes, there are 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 (cf. Falk & Szech, 2013). Imagining what the reasonable person would have done is a common procedure in the law (Green, 1967; Lagnado & Gerstenberg, 2017), and it has been argued to play an important role in how we assign responsibility (Fincham & Jaspars, 1983; Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, & Tenenbaum, 2014; **?**). Applying the CSM to domains outside of physics highlights the need to study what counterfactual contrasts people consider, and how they combine the outcomes of different contrasts to reach their judgment (Schlottmann, Allen, Linderoth, & Hesketh, 2002; Schlottmann, Ray, Mitchell, & Demetriou, 2006).

Conclusion

How do people make causal judgments? This paper presents a novel theory: the *counterfactual simulation model* (CSM) of causal judgment. 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, and 3) there are multiple aspects of causation which correspond to different ways of making a difference to the outcome. According to the CSM, people represent the world in terms of intuitive theories (Gerstenberg & Tenenbaum, 2017; Lake et al., 2016; Wellman & Gelman, 1992), and they make causal judgments by running counterfactual simulations (Gerstenberg, Peterson, et al., 2017; Kahneman & Tversky, 1982).

As a case study, we used the CSM to explain people's judgments about dynamic collision events. The CSM shows how people's causal judgments are influenced by different aspects of causation. These aspects capture the extent to which the candidate cause was necessary and sufficient for the outcome to occur, as well as whether it affected how the outcome actually came about. These aspects of causation provide a new conceptual landscape for analyzing the mapping between events in the world and the words we use to describe them. The CSM naturally handles judgments about situations in which something almost happened, or situations of omission in which nothing happened. It also provides a natural account of how people make judgments about physical support.

Some important challenges remain. Future versions of the CSM will need to incorporate the role of normative expectations, adequately handle situations of preemption, and capture people's causal judgments in richer domains that require more than just physical simulation. We will also need to think more about the relationship what role, if any, counterfactual simulation plays in causal perception, and what role causal judgments play in achieving our goals. These challenges notwithstanding, we believe that the CSM provides a powerful framework for asking these questions, and for thinking about different possible ways of answering them.

References

Abelson, R. P., & Kanouse, D. E. (1966). Subjective acceptance of verbal generalizations. In S. Feldman (Ed.), *Cognitive consistency: Motivational antecedents and behavioral consequents* (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` 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.

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.

Bechlivanidis, C., & Lagnado, D. A. (2013). Does the "why" tell us the "when"? *Psychological Science*, *24*(8), 1563–1572.

Bechlivanidis, C., & Lagnado, D. A. (2016, jan). Time reordered: Causal perception guides the interpretation of temporal order. *Cognition*, *146*, 58–66. Retrieved from `https://doi.org/10.1016%2Fj.cognition.2015.09.001` doi: 10.1016/j.cognition.2015.09.001

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The oxford handbook of causation*. Oxford University Press, USA.

Bello, P., & Khemlani, S. S. (2013). A model-based theory of omissive causation. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Bernstein, S. (2014). Omissions as possibilities. *Philosophical Studies*, *167*(1), 1–23.

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., 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.

Buss, A. R. (1978). Causes and reasons in attribution theory: A conceptual critique. *Journal of Personality and Social Psychology*, *36*(11), 1311–1321.

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.

Clarke, R., Shepherd, J., Stigall, J., Waller, R. R., & Zarpentine, C. (2015). Causation, norms, and omissions: A study of causal judgments. *Philosophical Psychology*, *28*(2), 279–293.

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.

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%2Fs10670-013-9439-2 doi: 10.1007/s10670-013-9439-2

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

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.

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%2Fxlm0000369 doi: 10.1037/xlm0000369

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

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (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., 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., & 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%2F0956797617713053` doi: 10.1177/0956797617713053

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. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2263–2268). Austin, TX: Cognitive Science Society.

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.

Girotto, 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.

Godfrey-Smith, P. (2010). Causal pluralism. In H. Beebee, 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., 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.

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

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. (1944). Social perception and phenomenal causality. *Psychological Review*, *51*(6), 358–374.

Heider, F. (1958). *The psychology of interpersonal relations.* John Wiley & Sons Inc.

Henne, P., Pinillos, Á., & De Brigard, F. (2016). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, 1–14.

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.

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

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2015). Not so innocent: Toddlers' inferences about costs and culpability. *Psychological Science*, *26*(5), 633–640. Retrieved from http://dx.doi.org/10.1177/0956797615572806 doi: 10.1177/0956797615572806

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.

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.

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.

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

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., 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., 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 https://doi.org/10.1017%2Fs0140525x16001837 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. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lewis, D. (1986). Postscript C to 'Causation': (Insensitive causation). In *Philosophical papers* (Vol. 2). Oxford: 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.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700–734. Retrieved from `http://dx.doi.org/10.1037/a0039655` doi: 10.1037/a0039655

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: Gen-*

eral, *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). Indicators of causal agency in physical interactions: The role of the prior context. *Cognition*, *132*(3), 485-490.

Mayrhofer, R., & Waldmann, M. R. (2016). Causal agency and the perception of force. *Psychonomic Bulletin & Review*, *23*(3), 789–796. Retrieved from `https://doi.org/10.3758%2Fs13423-015-0960-y` doi: 10.3758/s13423-015-0960-y

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.

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.

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.

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162.

Paul, L. (1998). Keeping track of the time: Emending the counterfactual analysis of causation. *Analysis*, *58*(3), 191–198.

Paul, L. A. (1998, jan). Problems with late preemption. *Analysis*, *58*(1), 48–53. Retrieved from `https://doi.org/10.1093%2Fanalys%2F58.1.48` 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%2F2678392` 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.

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.

Rips, L. J., & Edwards, B. J. (2013, jan). Inference and explanation in counterfactual reasoning. *Cognitive Science*, *37*(6), 1107–1135. Retrieved from http://dx.doi.org/10.1111/cogs.12024 doi: 10.1111/cogs.12024

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

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.

Samland, J., & Waldmann, M. R. (2015). Highlighting the causal meaning of causal test questions in contexts of norm violations. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2092–2097). Austin, TX: Cognitive Science Society.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

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%2Fj.actpsy.2006.05.005 doi: 10.1016/j.actpsy.2006.05.005

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a bayesian model of theory of mind. *Current Opinion in Psychology*, *17*, 15 - 21. Retrieved from http://www.sciencedirect.com/science/article/pii/S2352250X17300283 (Emotion) doi: https://doi.org/10.1016/j.copsyc.2017.04.019

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., Allen, D., Linderoth, C., & Hesketh, S. (2002). Perceptual causality in children. *Child Development*, *73*(6), 1656–1677. Retrieved from `https://doi.org/10.1111%2F1467-8624.00497` doi: 10.1111/1467-8624.00497

Schlottmann, A., Ray, E. D., Mitchell, A., & Demetriou, N. (2006, sep). Perceived physical and social causality in animated motions: Spontaneous reports and ratings. *Acta Psychologica*, *123*(1-2), 112–143. Retrieved from `https://doi.org/10.1016%2Fj.actpsy.2006.05.006` doi: 10.1016/j.actpsy.2006.05.006

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., 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.

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

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

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. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 619–644). Oxford University Press.

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., & Waldmann, M. R. (2016). Answering causal queries about singular cases. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2795–2801). Austin, TX: Cognitive Science Society.

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.

Suppes, P. (1970). *A probabilistic theory of causation*. Amsterdam: North-Holland.

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

Tang, Z. (2015, feb). Absence causation and a liberal theory of causal explanation. *Australasian Journal of Philosophy*, *93*(4), 688–705. Retrieved from https://doi.org/10.1080%2F00048402.2014.1001993 doi: 10.1080/00048402.2014.1001993

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.

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., 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., 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).

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018, April). Stable Causal Relationships Are Better Causal Relationships. *Cognitive Science*. doi: 10.1111/cogs.12605

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4 (ISBN 0-387-95457-0)

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014, Jan). One and done? optimal decisions from very few samples. *Cogn Sci*, *38*(4), 599-637. Retrieved from http://dx.doi.org/10.1111/cogs.12101 doi: 10.1111/cogs.12101

Waldmann, M. R., & Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216–227.

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. (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. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–147. Retrieved from http://dx.doi.org/10.1037/0033-295x.113.1.132 doi: 10.1037/0033-295x.113.1.132

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*(3), 580–601.

White, P. A. (2011). Visual impressions of force exerted by one object on another when the objects do not come into contact. *Visual Cognition*, *19*(3), 340–366. Retrieved

from https://doi.org/10.1080%2F13506285.2010.532379 doi: 10.1080/13506285.2010.532379

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

White, P. A. (2017). Visual impressions of causality. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning.* Oxford University Press.

Willemsen, P. (2016). Omissions and expectations: A new approach to the things we failed to do. *Synthese*. Advance online publication. doi: 10.1007/s11229-016-1284-9

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). The problem of variable choice. *Synthese*.

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.

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

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 participant judgments.*

Clip	Ball	Outcome				Cause				Model				Rating
		both	only A	only B	neither	difference	whether	how	sufficient	CSM_W	CSM_{WH}	CSM_{WHS}	Heuristic	
1	A	0	0	0	0	1	31	1	34	29	51	51	61	42
1	B					1	30	1	6	28	50	43	52	37
2	A	0	0	0	0	0	0	0	0	0	0	0	38	21
2	B					0	0	0	0	0	0	0	36	19
3	A	1	0	0	0	1	85	1	4	79	76	67	68	76
3	B					1	78	1	14	73	72	67	68	75
4	A	1	0	0	0	1	83	1	30	77	75	73	57	63
4	B					1	92	1	6	86	79	71	53	78
5	A	0	0	1	0	1	72	1	8	67	70	63	75	71
5	B					1	3	1	10	2	38	32	57	22
6	A	0	0	1	0	1	70	1	4	66	69	61	54	73
6	B					1	16	1	17	15	44	39	48	22
7	A	1	0	1	0	1	15	1	10	14	43	37	67	59
7	B					1	80	1	92	75	73	88	79	79
8	A	1	0	1	0	0	0	0	0	0	0	0	34	7
8	B					1	63	1	87	59	65	79	79	92
9	A	0	1	0	0	0	0	0	0	0	0	0	12	8
9	B					1	90	1	14	84	78	72	89	90
10	A	0	1	0	0	1	31	0	6	29	14	16	10	23
10	B					1	62	0	9	58	29	30	42	55
11	A	1	1	0	0	1	80	1	73	75	73	83	80	93
11	B					0	0	0	0	0	0	0	4	4
12	A	1	1	0	0	1	73	1	62	68	70	77	49	77
12	B					1	6	1	4	6	39	32	49	37
13	A	0	1	1	0	1	27	0	12	25	12	15	6	8
13	B					1	81	0	11	75	37	39	47	64
14	A	0	1	1	0	1	47	0	19	43	22	26	20	22
14	B					1	29	0	22	27	13	19	14	18
15	A	1	1	1	0	1	16	1	89	15	44	58	59	76
15	B					1	13	1	88	13	43	56	59	76
16	A	1	1	1	0	1	21	1	98	19	46	62	80	92
16	B					0	0	0	0	0	0	0	11	4
17	A	0	0	0	1	1	32	1	87	30	51	65	71	69
17	B					1	24	1	90	22	47	62	54	46
18	A	0	0	0	1	1	40	1	89	37	55	69	49	63
18	B					1	22	1	82	20	47	59	45	66
19	A	1	0	0	1	1	72	1	11	68	70	64	55	53
19	B					1	73	1	10	68	70	63	55	49
20	A	1	0	0	1	1	62	1	19	58	65	61	53	41
20	B					1	76	1	16	71	72	67	56	71
21	A	0	0	1	1	1	60	1	77	56	64	75	60	80
21	B					1	13	1	8	12	43	36	56	18
22	A	0	0	1	1	1	84	1	93	78	75	90	57	60
22	B					1	65	1	22	61	67	63	50	42
23	A	1	0	1	1	0	0	0	0	0	0	0	-8	3
23	B					1	74	0	7	69	34	35	31	39
24	A	1	0	1	1	1	56	1	12	53	62	57	50	44
24	B					1	97	1	5	91	81	73	52	73
25	A	0	1	0	1	1	52	1	2	48	60	52	60	43

Clip	Ball	Outcome				Cause				Model				Rating
		both	only A	only B	neither	difference	whether	how	sufficient	CSM_W	CSM_{WH}	CSM_{WHS}	Heuristic	
25	B					1	88	1	95	82	77	92	72	73
26	A	0	1	0	1	1	46	1	7	43	58	51	64	39
26	B					1	86	1	94	80	76	91	61	69
27	A	1	1	0	1	1	66	1	13	62	67	61	72	80
27	B					0	0	0	0	0	0	0	10	6
28	A	1	1	0	1	1	71	1	31	67	69	68	85	89
28	B					0	0	0	0	0	0	0	-5	5
29	A	0	1	1	1	1	52	1	4	48	60	52	58	47
29	B					1	74	1	9	69	70	64	63	67
30	A	0	1	1	1	1	69	1	2	65	68	60	60	58
30	B					1	84	1	8	78	75	68	66	56
31	A	1	1	1	1	1	6	1	2	5	39	31	53	44
31	B					1	6	1	4	5	39	31	58	46
32	A	1	1	1	1	0	0	0	0	0	0	0	8	5
32	B					1	49	1	11	46	59	53	67	71