Cognition 216 (2021) 104842

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A counterfactual simulation model of causation by omission

Tobias Gerstenberg [a,*], Simon Stephan [b]

[a] Stanford University, USA
[b] Göttingen University, Germany

ARTICLE INFO

Keywords:
Causation
Omission
Counterfactuals
Mental simulation
Intuitive physics

ABSTRACT

When do people say that an event that did not happen was a cause? We extend the counterfactual simulation model (CSM) of causal judgment (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021) and test it in a series of three experiments that look at people's causal judgments about omissions in dynamic physical interactions. The problem of omissive causation highlights a series of questions that need to be answered in order to give an adequate causal explanation of why something happened: what are the relevant variables, what are their possible values, how are putative causal relationships evaluated, and how is the causal responsibility for an outcome attributed to multiple causes? The CSM predicts that people make causal judgments about omissions in physical interactions by using their intuitive understanding of physics to mentally simulate what would have happened in relevant counterfactual situations. Prior work has argued that normative expectations affect judgments of omissive causation. Here we suggest a concrete mechanism of how this happens: expectations affect what counterfactuals people consider, and the more certain people are that the counterfactual outcome would have been different from what actually happened, the more causal they judge the omission to be. Our experiments show that both the structure of the physical situation as well as expectations about what will happen affect people's judgments.

1. Introduction

Suzy is on vacation and her friend Billy agreed to water her plants while she is away. When Suzy returns home, she is shocked to find out that all her plants have died. Billy forgot to water them! The verdict is clear: The plants died because Billy did not water them. While this judgment feels intuitive it raises problems for theories of causation. This scenario – which is familiar to causal enthusiasts – illustrates the problem of causation by omission. The outcome happened because Billy *didn't do* something. However, if we allow for non-actions (or, more broadly, non-events) to be causes, how can we curb the incoming onslaught of other omissive causes? For example, did not the plants also die because the Queen of England did not water them, or because the fire alarm sprinkler system did not go off?

Omissions have a complicated causal status in philosophy (Beebee, 2004; Bernstein, 2014, 2015; Hall, 2004; Lewis, 2004; Menzies, 2006; Schaffer, 2000). There are two major philosophical frameworks for thinking about causation: dependence theories and process theories. Counterfactual theories of causation, which belong to the first class of theories, analyze causal relationships in terms of counterfactual dependence (Lewis, 1973). According to these theories *c* caused *e* when both *c* and *e* happened, and when *e* would not have happened had *c* not happened. Counterfactual theories need not draw a fundamental distinction between omissive causation (when an outcome of interest happened as a consequence of the absence of an event) and commissive causation (when the outcome resulted from an event). Billy's not watering caused the plants to die because had he watered them, they would have survived. To accommodate the intuition that Billy was causally responsible but not the Queen of England, counterfactual theories have incorporated normative considerations (Halpern & Hitchcock, 2015; Hitchcock & Knobe, 2009). Expectations about what normally happens, or about what should happen, influence what counterfactuals are considered, and what counterfactuals are considered subsequently affects causal judgments (Kominsky & Phillips, 2019). The plants died because of Billy's not watering them, rather than the Queen's not watering them, because Billy was expected to water them whereas the Queen was not.

Process theories of causation (e.g. Dowe, 2000; Fair, 1979; Salmon, 1994) analyze causal relationships in terms of spatio-temporally contiguous transmission of a conserved quantity. Here, *c* caused *e*

* Corresponding author at: Stanford University, Department of Psychology, 450 Jane Stanford Way (Building 420), Office 302, Stanford, CA 94305, USA.
E-mail address: gerstenberg@stanford.edu (T. Gerstenberg).

https://doi.org/10.1016/j.cognition.2021.104842
Received 12 June 2020; Received in revised form 2 July 2021; Accepted 6 July 2021
Available online 21 July 2021

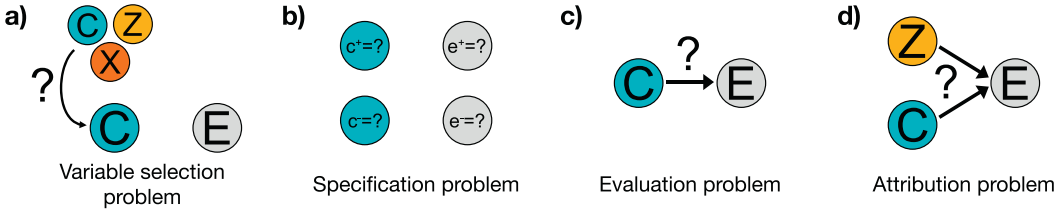

Fig. 1. The problem of giving a causal explanation for why e^+ happened can be broken down into four sub-problems. a) The *variable selection problem* is about what variables to consider as possible causes. Here, variable C was selected whereas X and Z were considered irrelevant. b) The *specification problem* is about what values the variables of interest can take on. Here, both variables are binary encoding whether an event happened (+), or did not happen (−). c) The *evaluation problem* is about how to assess whether there is a causal relationship between the candidate variables. This relationship could be established, for example, by considering the counterfactual of whether e^+ would have happened if c^+ had not happened. d) The *attribution problem* is about how to determine the extent to which each variable is causally responsible for the outcome. In a situation in which both C and Z caused E, the question is how much responsibility each candidate cause bears for the outcome.

when *c* transferred some quantity, such as physical force, to *e*. Traditionally, process theories have had trouble accommodating omissive causation because no force is being transmitted from the putative cause to the effect. However, the *force dynamics model* (Wolff, 2007) – a psychological process theory of causation inspired by linguistic research (Talmy, 1988) – handles omissive causation (Wolff, Barbey, & Hausknecht, 2010). It deems absences as causal when they correspond to the removal of a force. That way, the force dynamics model restricts the scope of possible causes to those that either have exerted a force on the effect in the past, or that were expected to do so.

Irrespective of which causal framework is used, there are a number of decisions that a causal modeler has to make in order to provide an adequate explanation for why a particular outcome happened. We have identified four sub-problems that need to be addressed (see Fig. 1). First, a modeler has to select what variables are relevant ("variable selection problem"). For any given outcome, there are typically a multitude of factors that may have contributed to that outcome. Furthermore, once omissions are allowed to be causes, there is a serious problem of causal proliferation – how should one decide what is causally relevant and what is not? Second, a modeler has to specify what values the variables of interest can take ("specification problem"). When saying that the plants died because Billy did not water them, we intuitively have the relevant contrasts in mind: namely, that the flowers would have survived if Billy had watered them. However, specifying the relevant contrast is often not trivial and will affect what causal verdicts are reached (see Schaffer, 2005, 2010). Third, there is the problem of evaluating whether, and if so how, the candidate causes affected the outcome ("evaluation problem"). While it has been proposed that causal relationships are assessed via mentally simulating what would have happened in relevant counterfactual situations, most accounts have not spelled out what this process might actually look like. Finally, if more than one variable has been identified as a cause of the outcome, to what extent should the outcome be attributed to each of the candidate causes ("attribution problem")? While all of these problems also arise for commissive causation, they are brought into greater relief when considering omissive causation as we will see below.

In this paper, we extend the *counterfactual simulation model* (CSM) of causal judgment to deal with omissive causation (Gerstenberg et al., 2021). The CSM predicts that people make causal judgments about physical events by mentally simulating what would have happened in relevant counterfactual situations. The model combines insights from both counterfactual and process theories of causation. From counterfactual theories, it incorporates the idea that causal judgments can be modeled by considering counterfactual simulations operating over a causal model of the situation. While counterfactual theories have traditionally modeled situations at a coarse level of granularity (e.g. using binary variables to represent the presence versus absence of events), the CSM draws insights from process theories that emphasize the fine-grained processes by which causes bring about effects (e.g. the physical laws that dictate how collision events play out). Gerstenberg et al. (2021) showed that the CSM accurately predicts people's causal judgments in a series of experiments that involve physically realistic dynamic collision events.

In our experiments, we show that people's causal judgments about omissions are influenced by their expectations, and that this effect of expectations on causal judgments is explained by assuming that expectations influence what counterfactual simulations people consider. We provide a concrete mechanism of how expectations affect what counterfactual situations people simulate, and show that this account is consistent with people's judgments.

The paper is organized as follows: We will first discuss the four sub-problems any comprehensive model of causal explanation needs to solve: (1) the variable selection problem, (2) the specification problem, (3) the evaluation problem, and (4) the attribution problem. We will then review two existing theoretical frameworks of how people make causal judgments about omissions. Subsequently, we will describe the counterfactual simulation model of causal judgment and its extensions to omissive causation. We tested the model's predictions in three experiments that address a subset of the four problems. Experiments 1 and 2 address the evaluation problem. Experiment 3 additionally addresses the specification and attribution problem by looking at a situation with multiple candidate causes. We conclude by discussing limitations of the model, and by pointing out promising avenues for future research.

1.1. Modeling causal explanations

To provide an adequate explanation of why something happened, reasoners need to address a series of sub-problems (see Fig. 1).[1] We will describe each problem in turn, taking the perspective of a causal modeler who aims to build an adequate representation of the situation. We assume that the causal modeler relies on the formal language of causal graphical models (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000). While we will describe each problem in turn, we do not mean to imply that there is a strict linear order in which the different problems are considered. In practice, the model building process is dynamic (see, e.g. Nyberg et al., 2021). Rather then stepping once through the four stages, a modeler is likely to cycle through the process several times, sometimes going back and forth between different sub-problems, until a satisfying model was generated.[2]

[1] These four sub-problems are not meant to be an exhaustive list of problems that an adequate account of causal explanations needs to address. Rather, they are problems that are brought into greater relief specifically when considering omissions as candidate causes. The distinction that we draw here between these problems may also not be quite as clear cut in practice. For example, the variable selection problem and the specification problem may be closely intertwined.

[2] We thank an anonymous reviewer for raising this point.

1.1.1. The variable selection problem

The first decision is what variables to include in the model (Fig. 1a). The variable selection problem has been discussed intensively in philosophy (e.g. Beebee, 2004; Bernstein, 2015; Hesslow, 1988; Willemsen, 2018). Only variables that are represented in the model are candidate causes of the outcome (Pearl, 2000). But what variables should be selected? For example, there are a large number of factors that may have contributed to Suzy's flowers dying (the flowers' need for water, the lack of water, the heat, …). The variable selection problem is severely aggravated once omissions are allowed as potential causes. Billy's not watering the flowers is a cause but so is the fact that the Queen of England did not water them. Allowing for omissive causation opens the flood gates to a proliferation of causes (cf. McGrath, 2005; Menzies, 2004; Wolff et al., 2010). The choice of variables will determine what causal conclusions are reached (Woodward, 2015), and constructing the right model is arguably often more an art than a science (Halpern & Hitchcock, 2011). So, how can we justify that Billy is included as a variable in the model, but not the Queen of England?

Psychologically speaking, the selection problem is the problem of what naturally comes to mind (Kahneman & Miller, 1986; Phillips, Morris, & Cushman, 2019) and, as it turns out, there are systematic factors that guide people's causal selections (Byrne, 2005; Girotto, Legrenzi, & Rizzo, 1991; Petrocelli, Percy, Sherman, & Tormala, 2011). The most prominent solution to the selection problem is to consider normative expectations (McGrath, 2005). What people deem causally relevant is influenced by what they expected to happen (or what should have happened): people tend to cite unexpected events as causes (Hart & Honoré, 1959/1985). Given the circumstances of the situation, Billy's not watering the flowers was unexpected, whereas the Queen's not watering them was expected – unless the Queen had been mentioned, one probably would have never even thought of her. Several studies have demonstrated how events that violate statistical norms (what used to happen in the past) or prescriptive norms (what should happen) are preferentially judged causes (Hitchcock & Knobe, 2009; Kahneman & Miller, 1986; Kahneman & Tversky, 1982; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; McGill & Tenbrunsel, 2000). Formal models of causation have been developed in which normative expectations influence which variables are deemed causes of the outcome (e.g. Gerstenberg et al., 2018; Hall, 2007; Halpern & Hitchcock, 2015).

While considering normative expectations helps, it does not fully solve the variable selection problem (see Livengood, 2011). We believe that a more general solution to this problem will require considering the purpose of inquiry. For example, we often construct causal models with the communicative purpose of providing an explanation to someone else (Achinstein, 1983; Hilton, 1990; Keil, 2006; Kirfel, Icard, & Gerstenberg, 2020; Potochnik, 2016; Turnbull & Slugoski, 1988). In that case, what variables are included in the model depends on what the speaker's assumptions are about what the listener already knows (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020). Often, the speaker's motivation will also influence variable selection and model construction more generally (Green, 2008; Kunda, 1987; Mercier & Sperber, 2011). For example, in legal arguments, the prosecution and defense will likely paint different pictures of what happened with the goal of constructing mental models in the jury members that lead them to find the defendent guilty or not (Fenton, Neil, & Lagnado, 2013). Note that at this stage, a person may also consider variables that are later concluded not to have affected the outcome. For example, a lack of nutrients in the soil may be considered a potential cause, only to conclude later that this did not cause the plants to die. A coroner might discover traces of poison in the victim's mouth and hypothesize that poisoning was the victim's cause of death. However, the autopsy might reveal that the poisoning was not responsible and that the actual cause of death was a heart attack that occurred before the poison could have entered into the victim's bloodstream (see Stephan, Mayrhofer, & Waldmann, 2020).

1.1.2. The specification problem

Once a modeler has decided what variables to include in the model, they have to determine what values these variables can take on. We call this the specification problem (Fig. 1b). Variables could be specified coarsely, for example, by simply having two possible values that represent whether or not an event happened. Variables could also be defined more finely, for example, by continuously specifying when and where the event happened (Gerstenberg et al., 2021; Lewis, 2000; Stephan et al., 2020).

Let's assume that we have selected variables C and E, and that each are specified as binary variables representing whether an event happened (+) or did not happen (−). For c^+ to have caused e^+, according to counterfactual theories of causation (Lewis, 1973; Paul & Hall, 2013), it needs to be the case that both c^+ and e^+ happened, and that e^+ would not have happened if c^+ had not happened.[3] To make this precise, a counterfactual theory needs to specify what the relevant events c^+ and e^+ are, and what it means for these events not to happen. As Schaffer (2005) puts it, counterfactual theories are inherently contrastive. Accordingly, the question of whether "C caused E" is a question about whether "c^+ rather than c^- caused e^+ rather than e^-". If Billy had watered the plants rather than not watered them, then the plants would have survived rather than died.

For positive events ("something happened"), the counterfactual contrast ("it didn't happen") is often well-defined. If Billy shot Steve (b^+) in the actual situation, then the counterfactual contrast of Billy not shooting Steve (b^-) is easy to imagine. However, when what actually "happened" was a negative event ("something didn't happen"), it is less clear what the relevant counterfactual contrast should look like. If Billy did not shoot Steve, how are we to imagine the event of Billy shooting Steve? Where would the shot have hit Steve? Would the bullet have gone straight through Steve's heart, or would the bullet merely have damaged some muscle fibers in Steve's arm?

This problem is further aggravated in models featuring multi-valued variables for which it is even less clear what the relevant counterfactual should be (see Hitchcock, 1995). For example, if a variable C can take on values 1, 2, or 3, the question of whether $C = 1$ caused e^+ is problematic because the counterfactual of what would have happened if $C = 1$ had not taken place is ambiguous. Maybe $C = 2$ would still have resulted in e^+, whereas $C = 3$ would have resulted in e^- (Halpern, 2016; Hitchcock, 1995; Lassiter, 2017; Livengood, 2011).

Moreover, when we consider variables that represent a person's action, then sometimes the relevant counterfactual contrast might not be one between acting and not acting, but rather between what the person did, and what someone else would have done in the same situation. The law often employs the reasonable person test in cases of negligence (i.e. when a person failed to act). The test asks whether a reasonable person would have behaved in a way such that the negative outcome would have been avoided (Green, 1967; Hart & Honoré, 1959/1985; Lagnado & Gerstenberg, 2017). Similar considerations also apply to other contexts such as evaluating sports performance (Gerstenberg et al., 2018). For example, when considering to what extent a basketball player is responsible for the team's success, we would not consider the counterfactual of what would have happened if the team had played 4 against 5. Rather, we might consider what would have happened if the player had been replaced with a substitute player.

Overall, just like for the selection problem, omissions render the specification problem more challenging. For positive events, there is often a single counterfactual contrast that comes to mind. In contrast, for negative events (i.e. events that did not happen), there are often multiple relevant counterfactual contrasts. The way in which the variables'

[3] While this definition of counterfactual dependence is overly simplistic and fails in situations in which the outcome is causally overdetermined (see, e.g. Halpern & Pearl, 2005; Lagnado, Gerstenberg, & Zultan, 2013), it is sufficient for our purposes here.

potential values are specified will affect what causal conclusions are reached. Judging whether C caused E depends on what counterfactual contrast is chosen, how this contrast is realized in one's mental simulation of the counterfactual, and one's evaluation of what the consequences would be.

1.1.3. The evaluation problem

Once the candidate variables have been selected and their potential values have been specified, the causal modeler needs to assess whether the variables of interest are actually causally related. Here, we focus on singular causal relationships ("Billy killed Steve.") rather than general causal relationships ("Weapons kill people.") (see also Danks, 2017; Stephan et al., 2020; Stephan & Waldmann, 2018). To reiterate, in order to determine whether the causal statement that "Steve died because Billy shot him." is true, one needs to evaluate whether Steve would not have died, if Billy had not shot him. Often, the evaluation problem is more challenging for omissive causation compared to commissive causation. To determine whether Steve survived because Billy did not shoot him, we need to evaluate whether a specified counterfactual contrast (e.g. a shot in the stomach) would have been deadly. Maybe the bullet would have missed all the vital organs? Maybe Steve would have been saved by an ambulance?[4] While the variable selection problem has received much attention in the context of causation by omission (e.g. Hesslow, 1988; Wolff et al., 2010), the evaluation problem has been largely neglected (but see Khemlani, Wasylyshyn, Briggs, & Bello, 2018, Experiment 4). The experiments we report below tackle the evaluation problem.

Petrocelli et al. (2011) suggested that the extent to which a certain counterfactual is relevant is a function of both how likely we are to consider it, and how likely it would have changed the outcome of interest. They propose a model in which a counterfactual's potency is determined by the product of an "if-likelihood" and a "then-likelihood". Thus, a counterfactual is potent only if both its if-likelihood and then-likelihood are high. Petrocelli et al. show across a number of experiments that counterfactual potency predicts judgments of regret, causation, and responsibility. For example, consider the counterfactual statement "IF Billy had watered the plants THEN they would have survived." The "if-likelihood" expresses the degree to which the antecedent condition of the counterfactual is perceived to be likely (i.e. that Billy had watered them). The "then-likelihood" can be thought of as capturing the conditional probability of the outcome as specified given that the antecedent condition is true (i.e. that the flowers would have survived). Here, both the "if-likelihood" and the "then-likelihood" are high, so this counterfactual is potent, and accordingly Billy's not watering the plants will be seen as a cause of the plant's death. In contrast, "If the Queen of England had watered the plants …" is not a potent counterfactual because the "if-likelihood" is low (while Billy was supposed to water the plants, the Queen had no obligation to do so).

While Petrocelli et al. note that the "if-likelihood" and "then-likelihood" are subjectively determined, they do not provide an account of how people generate these likelihoods (see also Wells & Gavanski, 1989).[5] We believe that people's intuitive understanding of the domain of interest determines both what counterfactuals come to mind as well as how to evaluate what would have happened in the relevant counterfactual situations. To make this proposal concrete, we will focus here on a relatively simple physical domain. Several accounts have linked judgments of causation to the relevance of counterfactuals, and that the relevance of counterfactuals may be determined via a sampling procedure (Icard, Kominsky, & Knobe, 2017; Kahneman & Tversky, 1982; Kominsky & Phillips, 2019). However, these accounts have not provided a concrete implementation of what this sampling process actually looks like. Below, we propose a model that does so.

1.1.4. The attribution problem

When multiple causes contributed to an outcome, how do people determine the extent to which each cause was responsible for the outcome? Even when the variable selection problem, the specification problem, and the evaluation problem have been solved, there is still the problem of attributing causal responsibility (Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Zultan, Gerstenberg, & Lagnado, 2012). Most of the psychological research on omissive causation has focused on this attribution problem (Bello & Khemlani, 2015; Clarke, Shepherd, Stigall, Waller, & Zarpentine, 2015; Henne, Bello, Khemlani, & Brigard, 2019; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Henne, Pinillos, & De Brigard, 2017; Khemlani et al., 2018; Livengood & Machery, 2007; Wolff et al., 2010; Wolff, Hausknecht, & Holmes, 2011).

In many studies, the candidate causes have been selected by the experimenters, the events of interest are clearly specified, and the causal relationships are easy to evaluate. These studies show that unexpected events that violate statistical or prescriptive norms tend to be judged as more causal (Hitchcock & Knobe, 2009; Kahneman & Miller, 1986; Kahneman & Tversky, 1982; Kominsky et al., 2015; McGill & Tenbrunsel, 2000; Sanna & Turley, 1996). For example, when two cars collide at an intersection, the driver who failed to brake at the red light is cited as the cause rather than the driver who failed to brake at the green light (see Clarke et al., 2015; Henne et al., 2017). In principle, either driver could have avoided the accident by stepping on the brakes, but it's normal for a driver to stop on a red light and to keep going when it's green. Formal models of causation have been developed that explicitly incorporate normative expectations (e.g. Gerstenberg et al., 2018; Hall, 2007; Halpern & Hitchcock, 2015).

Recent work has shown that in addition to normative expectations, the structure of the situation also affects causal attributions, leading people to sometimes prefer normal over abnormal events as causes (Gerstenberg & Icard, 2019; Harinen, 2017; Henne, Niemi, et al., 2019; Icard et al., 2017; Kirfel et al., 2020; Kominsky et al., 2015; Samland & Waldmann, 2016). Different factors influence people's expectations, including statistical information about past events, prospective norms about what behavior is appropriate in a given situation, as well as norms of proper functioning that take into account what function an artifact is supposed to fulfill (Hitchcock & Knobe, 2009). Expectations also influence moral judgments, such as attributions of blame. Generally, we blame people more for failure when we expected them to succeed (Gerstenberg, Ejova, & Lagnado, 2011; Gerstenberg et al., 2018). Even young children evaluate a person refusing to help more negatively, when it would have been easy for the person to help (Jara-Ettinger, Tenenbaum, & Schulz, 2015).

1.2. Existing theories of omissive causation

We have argued that in order to provide an adequate causal explanation of what happened, a modeler has to address a number of challenges. They have to select relevant variables, specify their possible values, evaluate the causal relationship between candidate causes and the outcome, and attribute the extent to which each variable was causally responsible for the outcome. These problems are aggravated for omissive causes. Currently, no computational model exists that addresses all of these problems. We will now discuss two existing theoretical frameworks for modeling causal judgments that have been extended to deal with causation by omission: the *force dynamics model*,

[4] Sometimes evaluating commissive causation may be more challenging. For example, considering what would have happened if a professor had not made it to class may be more difficult than imagining what would have happened if they had made it. We thank an anonymous reviewer for this suggestion.

[5] It should also be noted that it matters how these likelihoods are computed. Instead of computing the conditional probability $p(e|c)$ which, when used as a guide for causal relationships could, for example, lead to the false conclusion that two effects of a common cause are directly causally related (see Hagmayer & Sloman, 2009), one should compute the interventional probability $p(e|\text{do}(c))$ which is sensitive to the causal relationships of the variables of interest (see Pearl, 2000).

and the *mental model theory*. Afterwards, we will describe our model, the counterfactual simulation model of omissive causation.

1.2.1. A force dynamics theory of omissive causation

According to the force dynamics model (Wolff, 2007; Wolff et al., 2010, 2011), causation is characterized as an interaction between an agent and a patient that involves a transfer of force. Different causal expressions such as "cause", "enable", and "prevent" map onto different configurations of force transfer that are formalized using vector calculus. For example, the force vector representations that map onto "caused" and "enabled" differ with respect to whether or not the agent and patient force vectors at the time of interaction point into the same direction (see Fig. 2).

While the original model was developed to handle commissive causation (Wolff, 2007), Wolff et al. (2010) extended the force dynamics model to explain cases of omissive causation. The general idea behind this extended model is that causation by omission is linked to the removal of an actual (or anticipated) force that previously prevented (or would have prevented) the outcome from occurring. By linking omissive causation to force removal, the force dynamics model addresses the variable selection problem. Only those variables are potentially relevant that encode an actual or anticipated removal of force. To make this more concrete, imagine a person pushing aside a jack that is holding up a car, whereupon the car falls to the ground. Here, the removal of the jack caused the car to fall to the ground. According to the force dynamics model, causation by omission is embedded within a double prevention relationship (cf. Collins, 2000; Dowe, 2001; Hall, 2004; Schaffer, 2000). The removal of the jack prevented the *actual force* that had previously prevented the car from falling down.

To illustrate a scenario in which an omission involves the removal of an *anticipated force*, consider an auto racing situation in which car A is headed toward the finishing line but its path is currently blocked by car B standing on the line. Just in time, the breakdown service car pulls car B with a rope and frees the way for car A to cross the line. Here, it seems appropriate to say that the service car allowed car A to cross the line.[6] If the service car had not pulled car B out of the way, racing cars A and B would have collided. So, in this case, the service car removed the anticipated force that car B would have had on car A, thereby allowing car A to cross the line. Wolff et al. (2010) tested their model in a series of experiments depicting video clips of scenes like this one, and found that the force dynamics model correctly captures participants' modal responses of which expression best captures what happened in each clip.

In line with the counterfactual simulation model that we present below, (Wolff et al., 2010, p. 193–194) argue "that people are able to conduct partial 'reenactments' of the processes that join forces in the world. A reenactment involves specifying the objects and the forces acting on those objects in a situation. It also involves carrying out a simulation showing what happens as a consequence of the forces acting on the objects. Causal reasoning is assumed to consist of such reenactments." Through the notion of an anticipated force, the force dynamics model incorporates some of the machinery from counterfactual theories of causation. An anticipated force is a force that would have affected the outcome had things turned out differently.

Grounding causation in physical forces helps restrict the set of candidate causes. However, it also presents a challenge for generalizing this causal analysis to domains that are not well characterized by physical forces. For example, it's not easy to see how Billy's not watering Suzy's plants removed a force. Wolff et al. (2010) discuss cases like these and concur that if their analysis were to be applied to cases like these, it would also struggle with the variable selection problem having to rely on notions of normality to determine which "figurative" forces are anticipated (Billy) and which ones are not (the Queen).

1.2.2. A mental model theory of omissive causation

Another psychological theory of causal judgment that has been extended to handle omissive causation is the mental model theory (Goldvarg & Johnson-Laird, 2001; Khemlani et al., 2018). According to the mental model theory, people reason causally by representing mental models as sets of possibilities. Different causal terms such as "cause" and "enable" map onto different sets of possibilities. "C causes E" means that given C, E occurs. "C enables E" means that given C, it is possible for E to occur (Khemlani, Barbey, & Johnson-Laird, 2014). More precisely, the logical possibilities that define that "C causes E" are $C \wedge E$, $\neg C \wedge E$, and $\neg C \wedge \neg E$. The possibility $C \wedge \neg E$ would not be consistent with "C causes E". In contrast, the possibilities that define "C enabled E" are $C \wedge E$, $C \wedge \neg E$, and $\neg C \wedge \neg E$, whereas the possibility $\neg C \wedge E$ is inconsistent. The theory predicts people's reasoning errors about causal relationships based on a tendency to consider some possibilities but not others (see also Khemlani, Bello, Briggs, & Harner, 2020). For example, upon hearing that "C causes E" people first consider the possibility that both C and E happened, and they may fail to subsequently consider further possibilities that are consistent with the stated general causal relationship (e.g. that E happened although C did not happen).

The mental model theory assumes that people make causal judgments by building and inspecting simulated causal relations (Khemlani et al., 2018). Whereas the force dynamics theory represents simulations via interacting force vectors, simulations in the mental model theory are represented as discrete possibilities. Within this framework, omissions are modeled by negating antecedent events. For example, the absence of A causes E ("Not A causes E") is consistent with the following possibilities $\neg A \wedge E$, $A \wedge E$, and $A \wedge \neg E$, but inconsistent with $\neg A \wedge \neg E$.

Khemlani et al. (2018) tested the predictions of their model in a series of experiments. In Experiments 1 and 2, participants were provided with causal statements and asked to indicate which possibilities were consistent with the statement. In Experiment 3, participants were given both a causal statement ("The lack of wind will cause the fire to dissipate.") and an assertion about what happened ("There is wind and the fire does not dissipate."), and they were asked to say whether both can be true. The results of these experiments show that people's selection of consistent possibilities (Experiments 1 and 2) and their truth judgments (Experiment 3) are well accounted for by mental model theory.

In Experiment 4, which is closest to the experiments we report in this paper, participants viewed physical animations of a ball heading toward an entrance of a tube with a Y-shaped exit. The scene also featured a movable gate next to the tube's entrance, and a goal whose position varied between animations. Participants viewed sets of three animations after which they were asked to select one out of three statements that only differed in their causal verb (e.g. "Not closing the gate caused/enabled/prevented the ball to score/from scoring."). Participants' selections were largely consistent with the predictions of the mental model theory, distinguishing between "omissive causation" and "omissive enabling". However, participants were generally reluctant to saying that an omission (not opening the gate, or not closing the gate) "caused" the outcome to happen.

2. A counterfactual simulation model of causation by omission

The counterfactual simulation model (CSM) of causal judgment (see Gerstenberg et al., 2021) assumes that people make causal judgments about physical events by comparing what actually happened with the outcome of mentally simulating relevant counterfactual situations (see also Sloman, Barbey, & Hotaling, 2009). In line with Wolff et al.'s (2010) force dynamics model, the model assumes that "people simulate the processes that produce causal relationships rather than simply specifying the dependencies that hold between one event or state and

[6] Interestingly, it does not feel right to say that the service car caused car A to cross the line. We will return to the question of how to explain differences between causal expressions such as "caused", "enabled", or "allowed" in the General Discussion.

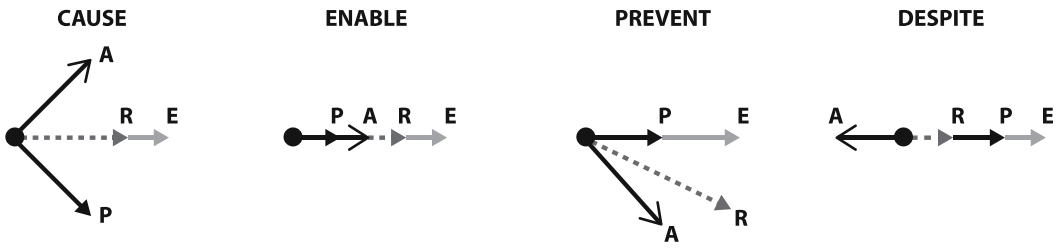

Fig. 2. According to Wolff's (2007) force dynamics model, causal expressions map onto configurations of force vectors. P = patient force, A = agent force, R = resulting force, E = endstate. For example, the construction "A caused P to reach E" maps onto a configuration in which P's initial force did not point toward the endstate E, and A's force combined with P's such that the resulting force R led P to reach the endstate. In contrast, the construction "A enabled P to reach E" implies that P's force vector already pointed toward the endstate, and A's force combined with P's such that it reached the endstate.

another." (p. 215) So far, the CSM has been successfully applied to capturing participants' judgments about physical events that actually happened (Gerstenberg et al., 2021). For example, Gerstenberg, Peterson, Goodman, Lagnado, and Tenenbaum (2017) have shown that there is a close mapping between people's causal judgments, and their subjective degree of belief that the outcome would have been different if the cause had not been present (see also Gerstenberg & Tenenbaum, 2016). Here, we extend the CSM to handle omissive causation by stipulating a concrete mechanism for how expectations affect the generation of counterfactuals. While our account does not address the selection problem, it does speak to the specification problem, evaluation problem, and attribution problem.

According to counterfactual theories of causation, causal claims are inherently contrastive and thus subject to the specification problem (cf. Schaffer, 2005). That is, when asking whether C caused E, what we are really asking is whether c rather than c' caused e instead of e'.[7] Formally, we define the probability that c caused e as

$$P(c \rightarrow e) = P(e'_{c'} | c, e). \quad (1)$$

Taking into account what actually happened (i.e. both c and e actually happened), we evaluate whether the alternative outcome e' would have happened, if C had been set to c' instead of its actual value c. To compute this probability, one needs to specify both what the relevant counterfactuals c' and e' are.

We borrow this formulation of the counterfactual directly from Pearl (2000), who developed a theory of causation that fits well within the interventionist tradition of counterfactual theories (cf. Woodward, 2003). According to interventionist theories, causal claims are analyzed by considering what the consequences of intervening in the putative causal event would have been. Roughly, c caused e if intervening on c would have made a difference to e (under the right circumstances).[8] This means that, in order to evaluate the causal relationship between c and e, one not only needs to consider what the relevant c' for a given c would have looked like, but also how c' would have come about. What intervention would have turned c into c' (cf. Gerstenberg, Bechlivanidis, & Lagnado, 2013; Lucas & Kemp, 2015)?

Consider the situation shown in Fig. 3a. Both Marble A and Marble B enter the scene from the right, collide with one another, and Marble B goes through the gate. Did Marble B go through the gate because Marble A hit it? In this case, specifying the relevant contrasts is fairly straightforward. Marble A's hitting Marble B (c) rather than not hitting it (c') caused Marble B to go through the gate (e) rather than miss the gate (e'). To generate a counterfactual in which Marble A had not hit Marble B, one can imagine an intervention which removed Marble A from the scene.[9] The CSM predicts that people will say that Marble B went through the gate because Marble A hit it. While an observer does not have direct access to what the outcome in the relevant counterfactual situation would have been, they can use their intuitive understanding of physics to simulate what path Marble B would have taken if Marble A had not hit it (see Gerstenberg et al., 2017, for eye-tracking evidence that this is in fact what people do). Here, it is clear that Marble B would have missed if the collision had not happened (as illustrated by the dashed path).

Compare this with the situation shown in Fig. 3b. Marble B enters the scene from the right and goes through the gate. Marble A remains stationary in the corner. Did Marble B go through the gate because Marble A *didn't* hit it? Put differently, did Marble A's not hitting Marble B (c) rather than hitting it (c') cause Marble B's going through the gate (e) rather than missing it (e')? Note how in the case of causation by omission, determining what c' should look like is less clear than in the case of causation by commission. While a relevant counterfactual readily comes to mind of what would have happened if Marble A had not hit Marble B (Fig. 3a), there are many relevant counterfactuals for what would have happened if Marble A had hit Marble B (Fig. 3b). One needs to intervene in the situation such that Marble A starts moving at some point in time, in some direction, with some velocity. In the given example, an observer has to first imagine one of the infinite possible situations in which Marble A had hit B, and then try to simulate what the consequences of that collision would have been.

The pair of situations shown in Fig. 3 is analogous to the scenario in which Billy shoots Steve that we used to illustrate the specification problem. When Billy shot Steve, it is relatively easy to imagine what would have happened if Billy had not shot Steve (cf. Fig. 3a). However, when Billy did not shoot Steve, it is less clear what would have happened if Billy had shot Steve (cf. Fig. 3b).

The evaluation problem is also more challenging here. Counterfactual simulations for omissions are more demanding than counterfactual simulations for physical events that actually happened. First, the relevant counterfactual antecedent (c') has to be generated. It's easy to imagine Marble A not hitting Marble B, whereas it's more challenging to imagine Marble A hitting Marble B. Second, the consequences of the counterfactual intervention (i.e. the transformation from c to c') have to be simulated. Again, in the case of causation by commission this involves simulating the trajectory that Marble B would have taken if Marble A had not hit it. It's relatively straightforward to mentally simulate how

[7] We use c and c' (as well as e and e') here instead of c^+ and c^-, because c could be a commission and c' an omission, or vice versa.

[8] The circumstances in which this counterfactual dependence holds may be different from what actually happened (see Halpern, 2016; Halpern & Pearl, 2005).

[9] Note that in its original formulation the CSM considers counterfactual operations on objects rather than events (Gerstenberg et al., 2021). That is, it imagines what would have happened if Marble A had not been present in the scene, rather than what would have happened if the collisions between Marble A and B had not happened. However, when applying the CSM to causation by omission, it's important to think about the relevant events that might have happened.

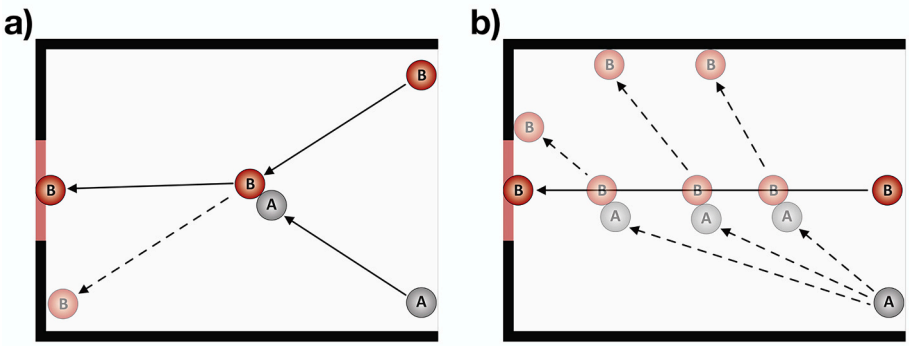

Fig. 3. Diagrammatic illustration of a) *causation by commission*, and b) *causation by omission*. In the *causation by commission* case, the question is whether Marble B went through the gate because Marble A collided with Marble B. Here, there is one relevant counterfactual that comes to mind – the counterfactual of what would have happened if the collision between Marble A and Marble B had not happened. In the *causation by omission* case, the question is whether Marble B went through the gate because Marble A *didn't collide* with Marble B. Both the specification problem as well as the evaluation problem are aggravated here compared to causation by commission. It is less clear what the relevant counterfactual is. There are many ways in which Marble A could have collided with Marble B that would have led to different counterfactual outcomes. And it is more difficult to evaluate what the consequences of Marble A colliding with Marble B would have been (the mental simulation involves reasoning about a collision event rather than merely extrapolating an unaffected marble's continuing motion).

Marble B's motion would have continued in that counterfactual situation (it just requires extrapolating the marble's observed motion path). In contrast, for causation by omission, one needs to simulate the consequences of Marble A's hitting Marble B. It's more challenging to mentally simulate how such a collision would have played out. For example, it's possible that Marble B would still have gone through the gate even if Marble A had hit it. Marble A could have only hit it lightly without affecting Marble B's path much. Or Marble A could have hit Marble B in such a way that it would have bounced off the top wall and gone into the gate regardless.

2.1. Scope of the model

Before laying out in detail how the CSM models omissions, we want to clarify the scope of the model. The CSM is a model of how people make causal judgments about *physical events* (see Gerstenberg et al., 2021). Accordingly, we restrict ourselves to settings in which participants are asked about omissions in dynamic physical interactions involving collision events, using scenarios like the one shown in Fig. 3. We will speculate in the General Discussion about how the model may be extended to handle omissions more broadly.

We further constrain ourselves to relatively simple situations for which it is feasible to implement a concrete model of the postulated simulation process. For example, in our setting, the relevant counterfactuals only involve simulating what would have happened if one of the objects in the scene had moved differently. We acknowledge that in many real world situations, evaluating the consequences of counterfactuals can be much more challenging, and people may lack the knowledge to accurately simulate how a counterfactual would have played out. By restricting ourselves to a simple domain, we can generate quantitative predictions from our model, and test these predictions experimentally.

The CSM does not provide a reductive account of causation in the philosophical sense: it does not reduce one concept (causality) to a more primitive one (counterfactual dependence, e.g. Lewis, 1973). Instead, our model builds on interventionist theories of causality (see Halpern, 2016; Pearl, 2000; Woodward, 2003). These theories begin with a causal representation of the domain, such as a system of structural equations that captures how the different variables in the model are causally related to one another. However, this general-level causal representation by itself does not yet yield answers to the question of what caused what to happen in a particular situation. To elucidate the concept of actual causation, these theories consider what the consequences of hypothetical interventions would have been. Here we build on this work by applying the same kind of machinery to modeling people's judgments about omissions in physical scenarios. Instead of using structural equations, we use a physics engine to express people's causal understanding of the domain (Gerstenberg & Tenenbaum, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). And instead of defining interventions as a change to the value of a variable, we implement interventions as operations on the objects in the physics simulation (e.g. making a ball move in a particular way).

There are many possible factors that influence what counterfactuals people bring to mind, including statistical norms ("what tends to happen") and prospective norms ("what should happen"). Much prior work has argued that causal judgments are influenced by what counterfactuals come to people's minds (Gerstenberg et al., 2021; Hilton & Slugoski, 1986; Icard et al., 2017; Kahneman & Miller, 1986; Kahneman & Tversky, 1982; Kominsky et al., 2015; Kominsky & Phillips, 2019). However, no work so far has tried to spell out concretely what this counterfactual simulation mechanism might look like. Here, we develop a concrete implementation of the idea that normative expectations influence what counterfactual possibilities people simulate, albeit in the restricted domain of simple physical interactions.

Our model predicts people's causal judgments about events that lie in the past (i.e. the model targets people's singular causation judgments, which concern the causes of events that have actually happened). In our everyday lives, causal judgments are often triggered by experiencing some unexpected or undesired outcome (see, e.g. Bohner, Bless, Schwarz, & Strack, 1988). Sometimes, we may also make prospective causal judgments about future events. For example, we may judge that not charging our phone will cause the battery to die. In this case, the relevant contrast is a future hypothetical rather than a counterfactual. While we do not look at these cases here, we believe that the same general principles for capturing people's causal judgments will hold. People consider hypothetical possibilities by imagining an intervention in the situation, and then simulating what the consequences of this intervention would be. Causal judgments about prospective events can be derived by simulating and then comparing two different hypothetical interventions: one in which the causal event of interest took place, and one in which it did not.

2.2. Modeling omissions

We assume that people make causal judgments about omissions by evaluating the outcome of counterfactual possibilities that are generated using their intuitive understanding of the situation (cf. Kahneman & Tversky, 1982). On this most general level, we believe that counterfactual theories of causation are applicable for handling omissions in a variety of different domains that include reasoning about physical events as well as psychological and social events (such as Billy forgetting to water the plants). People's intuitive understanding of the domain will dictate what counterfactuals come to mind (Billy should have watered the plants), and what would have happened in these counterfactual

situations (the plants would have survived). Here, we will illustrate how this account works by focusing on physical events. Specifically, we model people's causal judgments about dynamic collision events between marbles.

Consider the situation depicted in Fig. 4a. In the actual situation, Marble A did not move and Marble B went through the middle of the gate. Did Marble B go through the gate because Marble A did not hit it? To answer this question, the model simulates what would have happened if Marble A had collided with B. To do so, the model needs to determine the time t at which Marble A would have started to move, the direction d in which it would have moved, and the velocity v with which it would have moved. Once these parameters are set, the model simulates what would have happened. For many combinations of values for t, d, and v, Marble A would not have collided with Marble B. The model discards all such situations since it evaluates what would have happened if Marble A had *hit* Marble B. In other words, the model conditions on the truth of the counterfactual antecedent (i.e. situations in which the collision occurs). For each situation in which the two marbles collide, the model records what the outcome would have been – would B have missed the gate, or would it still have gone through the gate? The model then computes the probability that Marble A's not hitting Marble B was a cause of Marble B's going through the gate (cf. Eq. (1)) by calculating the proportion of samples in which B would have missed the gate instead of going through.

2.3. Expectations shape counterfactual simulations

An important question is what values t, d, and v should take on, which jointly determine what counterfactual situations are considered. The CSM generates counterfactual samples in two steps: a planning step and an implementation step. In the planning step, the CSM generates a set of "ideal path" counterfactual possibilities in which the causal event of interest happened, and the outcome would have been different from what actually happened. For example, Fig. 4a shows such an ideal path in which Marble A would have collided with Marble B, and Marble B would have missed the gate. In the implementation step, the CSM then samples one of the ideal paths and applies implementation noise to its execution. The figure shows two such sampled paths.

Let us illustrate these two steps with an analogy of a pool player who tries to knock one of the billiard balls into the pocket. We assume that the player is able to accurately plan the shot. For any given goal, there are multiple possible shots that would bring about that goal (e.g. striking the cue ball at different angles with different speed and spin, etc.). However, the actual shot is subject to some implementation noise (e.g. some degree of noise in the execution of the motor movements) that might lead to the goal not being accomplished. We would expect an expert player to be more likely to make a shot compared to a novice. In this setting, we would model the effect that expertise has on expectations by manipulating the degree of implementation noise. Whereas an expert has little implementation noise, a novice has more. Of course, experts and novices also differ in their ability to accurately plan their shot by mentally simulating the paths that the balls would take. For our purposes, we make the simplifying assumption that differences in expectations can be modeled via the degree of implementation noise.

Leaving the analogy, the CSM first generates a set of candidate counterfactual samples in the following way: it discretizes the space for the time t at which Marble A starts moving, the direction d in which it moves, and its velocity v. For t, the model considers all values from 0 to $t_{outcome}$, where 0 corresponds to the time at which Marble B starts moving and $t_{outcome}$ to the time at which Marble B went through the gate (or hit the wall). For d, the model considers the full range from Marble A going straight to the left to going straight up. For v, it considers a reasonable range from Marble A moving slowly to Marble A moving fast. For each generated world, the model notes whether Marble A and Marble B collided, and whether B went through the gate or missed the gate. It then discards all situations in which the two marbles did not collide. For the remaining situations, the model records whether Marble B would have gone through the gate or would have missed it.

To determine the probability that the outcome in the counterfactual samples would have been different from what actually happened, it uniformly samples one of the ideal paths generated in the first step and slightly perturbs the initial velocity vector that Marble A had in that sampled situation. We perturb both the magnitude and angle of Marble A's original velocity vector by adding Gaussian noise to the x-component and the y-component ($V'_x = V_x \cdot \mathcal{N}(1, \theta)$ and $V'_y = V_y \cdot \mathcal{N}(1, \theta)$). To compute $P(x \rightarrow y)$ (see Eq. (1)), the model then calculates the proportion of sampled paths for which the outcome would have been different from what actually happened.

We assume that expectations affect what counterfactuals people consider, and we capture the effect of expectations on the generation of counterfactuals through the θ parameter. For example, if an observer has strong expectations that Marble A will prevent Marble B from going through the gate (e.g. based on prior experience), we model this by only introducing a small amount of noise to Marble A's initial velocity vector (i.e. θ is small). If instead an observer does not have any strong expectations, we model this by introducing a larger degree of noise to Marble A's initial velocity vector (i.e. θ is larger).

3. Experiment 1: Expectations affect omissive causation judgments

Experiment 1 tests whether the CSM captures people's causal judgments for omissions in dynamic physical scenes. We look at causal judgments about situations in which Marble A failed to hit Marble B, and Marble B either went through the gate or missed it. Fig. 4 illustrates what the different clips looked like. In both clips, Marble A just rests still in the corner. In Fig. 4a Marble B goes through the gate. In Fig. 4b Marble B misses the gate. The CSM predicts that the extent to which people agree that the outcome happened because Marble A did not hit Marble B is a function of their degree of belief that the outcome would have been different in the relevant counterfactual. To test the hypothesis that counterfactual simulations map onto causal judgments, we ask one group of participants to indicate what they think would have happened if Marble A had hit Marble B, and another group of participants to judge whether the outcome happened because Marble A did not hit Marble B. Furthermore, we investigate how different types of expectations (statistical or social) influence people's judgments.

3.1. Methods

3.1.1. Participants

517 participants participated in this experiment. 47 participants who either failed to answer a simple attention or memory check question were excluded prior to any data analysis. The attention check question referred to a video clip subjects were shown on a separate screen towards the end of the study in which both marbles went through the gate. Subjects were asked to evaluate the statement "Both Marble A and Marble B went through the gate." (options: "True" versus "False"). The memory check query was presented after the attention check question. Subjects were asked to evaluate the statement "Marble A's color was grey, and Marble B's color was blue." (options: "True" versus "False" with "True" being the correct answer). The remaining 476 participants (251 female, 225 male, $M_{age} = 34$, $SD_{age} = 12$) who provided valid data received a monetary compensation of £ 0.25.

For all of the experiments reported here, participants were recruited via Prolific (www.prolific.co) using the following inclusion criteria: minimum age of 18 years, English as native language, an approval rate of at least 90 percent, and no previous participation in other studies of this project including pilot studies. We also asked participants to participate only via laptop or desktop computer and not via smartphone or tablet because we wanted to minimize the chances that participants take part who are in environments that might distract them (e.g. public

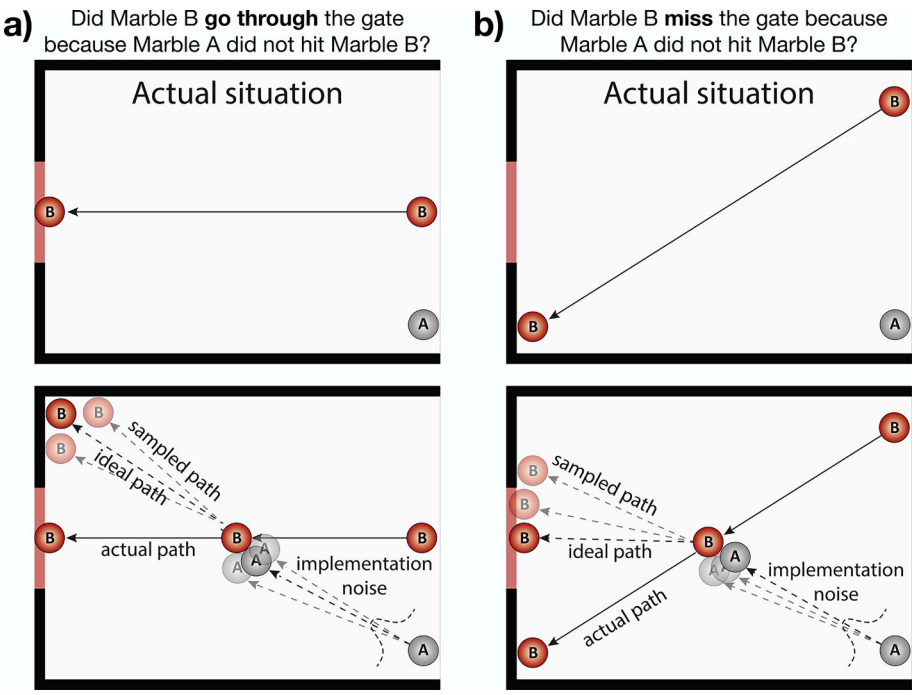

Fig. 4. Experiment 1 diagrams: Illustration of what actually happened (top) and the counterfactual simulation model (bottom). The diagrams illustrate the actual path that Marble B took, as well as an ideal path for a) Marble A preventing Marble B from going through the gate, or b) Marble A causing Marble B to go through the gate. The sampled paths show example simulations that result from applying implementation noise to the ideal path. *Note*: In a) Marble A would have prevented Marble B from going through the gate for both sampled paths. In b) Marble A would have caused Marble B go through the gate in one sample but not so in the other in which Marble B would still have missed even though Marble A hit Marble B.

places, subway).

3.1.2. Design, materials, and procedure

In the experiment, we manipulated information that affected participants' expectations about what will happen (*expectation*: no expectations, statistical expectation, social expectation) and what actually happened (*actual outcome*: missed versus went through). Finally, we varied whether participants answered a causal question or a probability question (*question*: causation versus probability). All factors were manipulated between participants. The probability question was: "What do you think are the chances that Marble B would have missed [gone through] the gate if Marble A had hit it?" Participants provided their responses on a continuous slider with the endpoints labeled "very unlikely" (0) and "very likely" (100). The causal test query was: "To what extent do you agree with the following statement: Marble B went through [missed] the gate, because Marble A did not hit Marble B." Participants responded on a continuous slider with the endpoints labeled "not at all" (0) and "very much" (100). The physical animations were created in Adobe Flash CS5 using the physics engine Box2D. All of the experimental materials, data, modeling, and analysis scripts are available online at https://github.com/cicl-stanford/omission.

3.1.3. No expectations condition

In the "no expectations" condition, participants simply read that they will see an animation featuring a stage with solid walls, two marbles, A and B, and a gate. All participants were shown a graphical illustration of the scene.

3.1.4. Statistical expectation condition

Participants in the "statistical expectation" condition were presented four primer clips in which Marble A hit Marble B. Participants who later saw the "went through" test clip, were shown four primer clips in which Marble A prevented Marble B from going through the gate. In each of these clips, Marble B would have gone through the gate if Marble A had not hit it. Participants who later saw the "missed" test clip, were shown four primer clips in which Marble A caused Marble B to go through the gate. In each of these clips, Marble B would have missed the gate if Marble A had not hit it. For each primer clip, participants indicated on a continuous slider to what extent they agreed with the statement "Marble B went through [missed] the gate, because Marble A hit Marble B." with the endpoints labeled "not at all" and "very much". Having participants provide judgments about these clips ensured that they were paying attention to the statistical expectation manipulation.

Based on what subjects observed in these primer clips, they should expect that Marble A would prevent Marble B from going through the gate for the "went through" test clip, and that Marble A would cause Marble B to go through the gate for the "missed" test clip.

3.1.5. Social expectation condition

In the "social expectation" condition, participants were instructed that the video clip will show what happened during a game of marbles played by two agents, Andy and Ben. Participants who later were shown the "went through" clip were told that it's Andy's job to hinder Ben's marble from going through the gate. Participants who later watched the "missed" clip were told that it's Andy's job to help Ben flip his marble through the gate. Based on this instruction, participants should expect that Andy will either prevent Ben's marble from going through, or help to knock it into the gate (depending on the outcome condition).

3.2. Model predictions

Experiment 1 manipulates prior information about what was likely to happen, as well as what actually ended up happening. The CSM predicts that people will generally agree more with the statement "Marble B went through the gate because Marble A didn't hit it." (Fig. 4a) than with the statement "Marble B missed the gate because Marble A didn't hit it." (Fig. 4b). Intuitively, this follows from the fact that, in this setting, hitting Marble B into the gate is more difficult than preventing Marble B from going through the gate. Put differently, a small degree of implementation noise added to a shot in which Marble A made Marble B miss the gate is still likely to lead to a miss (Fig. 4a bottom panel). In contrast, a small degree of implementation noise added to a shot in which Marble A made Marble B go through the gate might lead to Marble B missing the gate instead of going through (Fig. 4b bottom panel). In short, hitting Marble B into the gate is more sensitive to noise than preventing Marble B from going through the gate.

The CSM further predicts that expectations will affect agreement judgments. As discussed above, we model the effect of expectations

through the θ parameter which affects how much noise is added to Marble A's initial velocity vector. If Marble A is expected to prevent Marble B from going through the gate, this can be captured by generating counterfactual samples with little noise. If less noise is added, it means that more of the counterfactual samples will be successful in generating an outcome that's different from what actually happened. The model predicts that agreement judgments should generally be higher in the statistical and social expectation conditions compared to the no expectations condition.

To sum up, we can derive the following two hypotheses from the CSM about participants' causal judgments in Experiment 1:

Hypothesis 1.1. Agreement judgments that the outcome happened because of the omission will be higher when Marble B went through the gate compared to when Marble B missed the gate.[10]

Hypothesis 1.2. Agreement judgments will be higher to the extent that the omitted event was expected to happen (and expected to be successful in undoing the actual outcome).

3.3. Results

Fig. 5 shows participants' causation ratings (red), probability ratings (blue), as well as the predictions of the counterfactual simulation model (CSM; black). Table 1 shows the results of a 2 (*question*) × 3 (*condition*) × 2 (*outcome*) between-participants ANOVA. There was a significant effect of *outcome*: judgments were higher when Marble B went through the gate than when it missed the gate, $\Delta M = 16.36$, 95% CI [10.58, 22.13], $t(464) = 5.57$, $p < .001$. There was also a significant effect of *condition*: judgments were higher in statistical and social expectation conditions compared to the no expectations condition, $\Delta M = 19.34$, 95% CI [13.23, 25.44], $t(464) = 6.22$, $p < .001$. Judgments in the statistical expectation condition were significantly lower than in the statistical expectation condition, $\Delta M = -10.00$, 95% CI [− 18.49, − 1.52], $t(464) = -2.77$, $p = .016$. There was also a significant interaction between *question* and *condition*.

We fitted the θ parameter in the CSM to the data by minimizing the squared error between model predictions and averaged judgments.[11] The CSM correctly captures the difference in agreement ratings for both the causation and probability condition as a function of the outcome (Hypothesis 1.1). Judgments were higher for the "went through" clip compared to the "missed" clip. While the exact predictions of the model depend on the value of the θ parameter, the model predicts this qualitative pattern for all values of $\theta > 0$ (see Fig. A1 in the Appendix).

The CSM also captures that the agreement ratings are overall higher in the statistical and social expectation conditions compared to the no expectations condition (Hypothesis 1.2). It accounts for this pattern by assuming that what counterfactual simulations participants sample differs between the conditions. To account for the data, the model adjusts the θ parameter which determines the degree of noise that is added to Marble A's initial trajectory. Consistent with Hypothesis 1.2, the best-fitting parameter is smaller for the social expectation condition ($\theta = 0.04$), and the statistical expectation condition ($\theta = 0.08$), compared to the no expectations conditions ($\theta = 0.40$). The model does not predict the interaction effect between *question* and *condition*. As Fig. 5 illustrates, the model only makes one prediction for both question types, since it assumes a direct mapping from the counterfactual probabilities to the causal judgments.

3.4. Discussion

The results of Experiment 1 support the idea that people make causal judgments about omissions by mentally simulating whether the outcome would have been different in the relevant counterfactual situation. The CSM correctly predicted that people would be more willing to agree that Marble B went through the gate because Marble A did not hit it (Fig. 4a) than they would be to agree that Marble B missed the gate because Marble A did not hit it (Fig. 4b). This prediction arises from the assumption that it's easier to imagine that a collision between the marbles would have prevented Marble B from going through the gate versus knocking it into the gate.

The CSM also captures the fact that causality and probability judgments were higher when participants had formed expectations about what would happen compared to when they had no expectations. The CSM explains this difference by assuming that expectations affect what counterfactuals people consider. Specifically, the model predicts that participants are more likely to simulate counterfactuals in which the outcome would have been different from what actually happened when this counterfactual is consistent with their expectations.

In contrast to what the CSM predicted, the mapping between counterfactual probability judgments and participants' agreement that the outcome happened because of Marble A was not perfect. For example, in the social expectation condition, the difference in participants' judgments between the two clips is more pronounced for the causation question than the probability question. This suggests that there may be additional factors that influence participants' answers to the causation question that cannot be reduced to considerations of the relevant counterfactual. For example, in the social expectation condition, one may have considered that Player B's initial failure to get the marble on the right track (Fig. 4b) signals an exceptionally bad performance, and thus Player A's failure to correct this mistake is seen as less causally relevant.

Experiment 1 highlighted both the specification problem and the evaluation problem. Even though it is clear what type of event needs to be considered in the counterfactual (namely Marble A's hitting Marble B instead of sitting still), there are many ways in which this event could have been realized. A generative model of the situation is required to simulate the physical process by which the causal event of interest had come about as well as what the consequences of the collision between the marbles would have been. Furthermore, the problem of causal attribution did not arise in this scenario because there was only a single candidate cause. Although, as we mentioned above, in the social condition one could argue that player B carries some of the responsibility in the "missed" condition for having shot the marble poorly. We will return to the attribution problem in Experiment 3.

The CSM predicts that the difference between participants' judgments for the two test clips arises from the fact that Marble A's hitting Marble B would have been more likely to have made a difference to the outcome when Marble B went through the gate compared to when it missed the gate. However, it is also possible that people generally provide higher causal judgments when a positive outcome comes about as the result of an omission ("causation by omission") as compared to a negative outcome ("prevention by omission"). In Experiment 2, we investigated whether such a general asymmetry between omissive causation and omissive prevention exists, or whether, as predicted by the CSM, this difference disappears once the specification problem and the evaluation problem do not arise.

4. Experiment 2: No asymmetry between omissive causation and omissive prevention for well-specified contrasts

The goal of Experiment 2 was to rule out that the observed difference

[10] Note that this is not a general prediction that judgments for positive outcomes will be higher than judgments for negative outcomes. Instead, this prediction derives specifically from the fact that given the setup of the situation, it is more likely that a collision between Marble A and Marble B would prevent Marble B from going through the gate (see Fig. 4a), than it would cause Marble B to go through the gate (see Fig. 4b).

[11] Given the low number of data points ($n = 6$, the means for the probability and causality question in the three different expectation conditions), we refrain from reporting quantitative model fits.

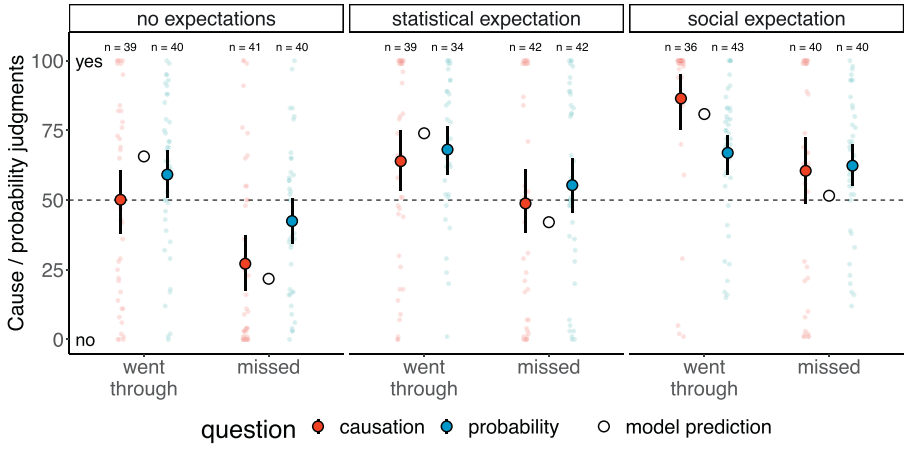

Fig. 5. Experiment 1 results: Agreement judgments for the "went through" clip and the "missed" clip (see top left and top right panel in Fig. 4, respectively). Large red dots indicate mean causation judgments, and large blue dots indicate mean probability judgments. Error bars are 95% bootstrapped confidence intervals. Small colored dots indicate individual responses (jittered along the *x*-axis for visibility). White dots indicate model predictions. Sample sizes are shown at the top. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Experiment 1 results: Main effects and interactions in a between-participants ANOVA with Question (probability, causality), Condition (no expectations, statistical expectations, social expectations), and Outcome (went through, missed) as predictors. *Note*: $\widehat{\eta}^2_G$ = generalized eta-squared, MSE = mean squared error.

Effect	F	df_1	p	$\widehat{\eta}^2_G$
Question	0.97	1	.326	.002
Condition	23.29	2	<.001	.091
Outcome	30.99	1	<.001	.063
Question × Condition	4.46	2	.012	.019
Question × Outcome	2.91	1	.089	.006
Condition × Outcome	0.36	2	.695	.002
Question × Condition × Outcome	0.97	2	.381	.004

$df_2 = 464$, $MSE = 1{,}023.20$.

between the "went through" and "missed" clips in Experiment 1 came about because people generally treat omissive causation differently from omissive prevention. The CSM only predicts a difference between two situations when the relevant counterfactual was more likely to be different in one situation compared to the other. Hence, our strategy in Experiment 2 was to hold this probability constant while manipulating the actual outcome. To achieve this goal, we simply replaced Marble A with a wall. To model the "missed" and "went through" situations in this setup, we varied whether or not the wall blocked the gate while the marble headed towards it (see the diagrams in Fig. 6). Participants rated how much they agreed that the marble went through the gate (or did not go through the gate) because the wall *did not move*.

4.1. Methods

4.1.1. Participants

65 participants (25 female, 40 male, $M_{age} = 33$, $SD_{age} = 13$) completed the experiment and received a monetary compensation of £ 0.25.

4.1.2. Design, materials, and procedure

The instructions were similar those used in the "no expectations" condition of Experiment 1. Participants saw an illustration that made it clear that the wall can only be in two different positions, either right in front of the gate or in the upper left corner of the stage and thus fully out of the way (see Fig. 6). We manipulated between participants whether the marble went through the gate or missed the gate (see Fig. 6). Participants viewed the test clip after having read the instructions, and then indicated on a continuous slider to what extent they agreed with the statement: "The marble went through the gate because the wall didn't move." or "The marble didn't go through the gate because the wall didn't move." depending on the outcome with the endpoints of the slider labeled "not at all" (0) and "very much" (100).

4.2. Model predictions

The CSM predicts that participants' causal judgments about omissions are determined by what they believe would have happened in the relevant counterfactual situation in which the event had taken place. For the situations depicted in Fig. 6 there is little to no uncertainty about the counterfactual outcome – it is clear that the outcome would have been different, had the wall moved (assuming that the wall would have moved fully out of the way before the marble arrived). Because it is clear that the outcome would have been different in the counterfactual situation from what actually happened, the CSM predicts that participants' causal judgments should be very high in both cases (and their should be no difference between the situation in which the marble went through the gate versus when it missed the gate).

Hypothesis 2.1. Agreement judgments that the outcome happened because the wall did not move (the omission) will be equally high for

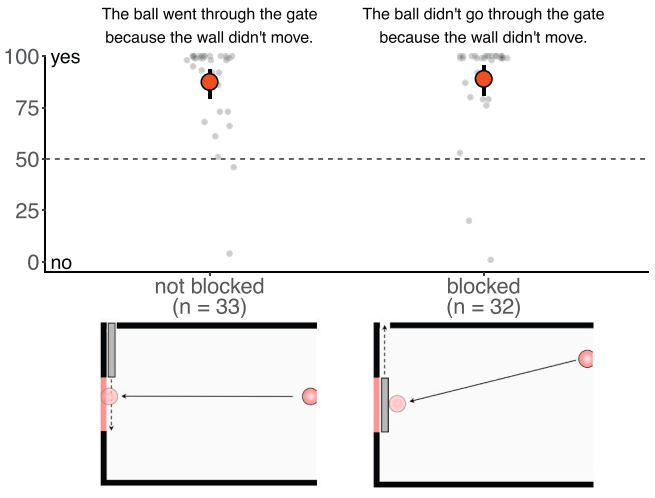

Fig. 6. Experiment 2 results: Participants' agreement ratings. In the "not blocked" condition (left), the wall was out of the way. In the "blocked" condition (right), the wall was in front of the gate. Large red dots indicate mean responses with 95% bootstrapped confidence intervals. Small black dots indicate individual responses (jittered along the *x*-axis for visibility). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

both the case in which the ball went through the gate, and in which the ball did not go through the gate. If there was a general tendency to treat omissive causation differently from omissive prevention, then participants' judgments should be greater for the clip in which the wall did not block the marble compared to the clip in which the wall blocked the marble.

4.3. Results

Fig. 6 shows participants' agreement judgments for the two clips. As predicted by the CSM (Hypothesis 2.1), participants gave very high causal ratings for both the "not blocked" clip ($M = 87.52$, $SD = 21.62$) as well as the "blocked" clip ($M = 89.00$, $SD = 23.21$). Participants' ratings did not differ significantly between the clips, $\Delta M = -1.48$, 95% CI [−12.60, 9.63], $t(63) = -0.27$, $p = .790$.

4.4. Discussion

The results of Experiment 2 suggest that there is no general asymmetry between judgments of causation by omission versus prevention by omission in the physical setting that we've explored. Instead, once it is clear what would have happened in the relevant counterfactual situation, participants agree that the outcome happened because the event of interest did not come about. We did not ask participants to evaluate the probability of the outcome in the counterfactual situation this time. However, participants' causal judgments are well explained by assuming they had little to no uncertainty about the counterfactual outcome. Whereas there were many possible ways in which Marble A could have hit Marble B in Experiment 1, in Experiment 2 the space of possibilities was drastically reduced such that the specification problem did not arise anymore. The wall can only be in two possible states (although there might still be some uncertainty about when it would move and how fast). The evaluation problem did not arise either since it's easy to mentally simulate what the outcome would have looked like had the wall moved instead of remaining still. The results also show that there is no general outcome effect on omissive causation judgments. For example, it's not the case that causal judgments for omissions are greater when the outcome was positive (or negative).

Together, the results of Experiment 1 and Experiment 2 support the view that causal judgments about omissions are sensitive to people's degree of belief about what would have happened in the relevant counterfactual situation. Both experiments have addressed the specification problem and the evaluation problem. Experiment 3 investigates participants' judgments in a setting that features multiple candidate causes. In this setting, the attribution problem arises of how much each candidate was responsible for the outcome.

5. Experiment 3: Attributing causal responsibility for multiple omissions

When a positive outcome did not happen because of several omissions, how do people decide which of the omissions was the most responsible? As alluded to in the introduction, a prominent suggestion for how to address this attribution problem is to consider the role of expectations (e.g. McGrath, 2005). Accordingly, we deem those omissions as causally relevant that we expected to happen. From the perspective of the counterfactual simulation model, expectations affect what counterfactuals come to mind (cf. Kahneman & Miller, 1986; Kahneman & Tversky, 1982). Hence, the idea is that omissions for which it is easier to imagine that they could have undone the outcome will be regarded as more causally relevant (cf. Gerstenberg et al., 2011; Petrocelli et al., 2011). Experiment 3 tests this prediction.

5.1. Methods

5.1.1. Participants

104 participants (49 female, 55 male, $M_{age} = 33$, $SD_{age} = 11$) participated in this pre-registered experiment (https://osf.io/fu9rq).

Concerning the sample size rationale, a pre-test with 62 participants yielded an effect of $d = 0.63$ for the observed mean difference of the blame ratings from the midpoint of the scale (in the expected direction). The observed effect for the difficulty ratings was even higher. With a sample of around 60 participants, an effect of $d = 0.63$ can be detected with more than a 99% probability using a directed one-sample t-test. We wanted to be more conservative and planned the main study with a smaller effect of $d = 0.4$, which has the power to be detected with 99% probability if a sample of $n = 100$ is tested. To have the same number of participants in each of the eight conditions (created by counterbalancing certain aspects of the test clip), we tested $n = 13$ participants per condition ($n = 104$ participants total).

5.1.2. Design, materials, and procedure

We instructed participants that they were going to watch a short video clip of a marbles game. Participants learned that the game was a team game in which two players ("Player Green" and "Player Blue") try to flip their marble such that they knock the gray stationary marble through the red gate. Fig. 7 shows a diagram illustrating one version of the test clip. Players take turns between rounds as to who flips their marble first and who second. In each round, the team gets a point if they manage to knock the gray marble into the gate. It does not matter which of the two players knocks it in. As Fig. 7 shows, there is a horizontal barrier next to one of the starting positions (for the green player's marble in this case). Players take turns as to who is assigned the starting position close to the barrier. We showed participants a picture of the playing field similar to the one in Fig. 7 but without the marble's motion paths.

After participants had read the instructions, they proceeded to the test phase. We asked two test questions in this study, a question about *difficulty*, and a question about *blame*. The question order was counterbalanced between subjects. Participants in the "blame first" condition were first shown a screen with the test video clip. They saw that both the green and the blue marble failed to hit the gray marble. The marbles' trajectories are shown in Fig. 7. We also counterbalanced between participants (1) whether the green or the blue marble moved first, (2) whether the green or the blue marble was positioned at the top, and (3) whether the horizontal barrier was positioned at the top or the bottom (Fig. 7 shows the configuration in which the barrier was at the top and the green marble moved first).

After participants had watched the clip, they were asked the blame test question, which was presented at the bottom of the same screen that showed the video clip. The question was introduced by the prompt "As you have seen, both the green and the blue player failed to hit the grey marble in this round. So, the team didn't score a point." The question was "Which player is more to blame that the grey marble did not go through the gate?". Ratings were provided on a sliding scale whose endpoints were labelled "Definitely the blue player" and "Definitely the green player". The midpoint was labelled "both equally". On the second screen of the test phase, participants were shown a stationary picture of the playing field with the following prompt: "Please have a look again at the picture below showing the configuration of the playing field and then answer the additional question below." They were then asked the difficulty question, which was "For which player is it more difficult to knock the gray marble into the gate?". Ratings were again provided on the same sliding scale. In the "difficulty first" condition, the order of presentation of the two screens was reversed. After participants provided their responses, they proceeded to a short demographic screen on which we also asked them for a brief explanation of their blame judgment.

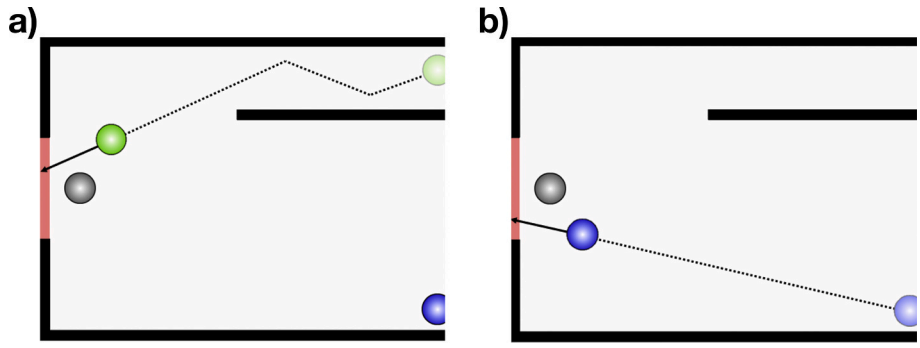

Fig. 7. Illustration of the animations used in Experiment 3. a) Green plays first and misses. b) Blue plays second and misses. The position of the barrier, the starting position of each player, and which player went first was counterbalanced between participants. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Model predictions

We generated predictions from the CSM by assuming that people mentally simulate how each player could have flipped their marble differently in order to knock the gray marble into the gate. Notice that in Experiments 1 and 2 the event of interest was clearly specified in the because statement (e.g. "Marble B went through the gate because marble A didn't hit it."). In contrast, in Experiment 3, we asked participants to evaluate whose players' shot was more difficult and who was more to blame for the negative outcome.

How the relevant counterfactual contrast is specified makes a difference here. In the actual situation, both the blue and the green marble missed the gray marble. If the counterfactual contrast was what would have happened if the green/blue marble *had hit* the gray marble, then there would almost be no difference between the two marbles. For both the green and the blue marble, the chances are high that the gray marble would have gone through the gate if it had been hit. However, if the counterfactual contrast is instead what would have happened if another (reasonably good) player *had shot* the marble instead, then there is a clear difference between the two. For the blue marble, it's easy to imagine how a player could have knocked the gray marble into the gate with a straight shot. For the green player, whose shot is obstructed by the obstacle, it's less clear how the marble would have needed to be shot such that it would have resulted in a successful outcome. So, how the counterfactual contrast is specified makes a big difference to the model's prediction.

To predict participants' judgments in Experiment 3, the model first constructs a space of counterfactual possibilities by considering in which direction each marble could have been shot. For Experiment 1, we needed to specify three parameters: the time at which the marble is shot, the direction in which it is shot, and the magnitude of the shot. This was necessary because the setting was dynamic. However, in this experiment the setting is static. All marbles are initially stationary. The relevant space of counterfactuals can therefore be constructed using a single parameter that captures the direction in which a marble is shot.[12]

To generate the space of counterfactuals, we used a finely discretized grid according to which each marble is flipped at an angle ranging from 100° to 260° (90° is north, and 180° is west). Like in Experiment 1, we then selected the subset of cases for which the counterfactual outcome of interest would have been realized (i.e. cases in which the gray marble was knocked into the gate). This space of possibilities includes situations in which the blue/green marble collide several times with the walls before knocking the gray marble into the gate. Intuitively, for the blue marble in Fig. 7b, a counterfactual comes to mind in which the marble would have directly knocked the gray marble into the gate (rather than colliding with one or several of the walls before). For the green marble, it is impossible to knock the gray marble into the gate without first colliding with the walls. The minimum number of wall collisions before knocking gray into the gate is two.[13]

To generate a counterfactual simulation, the model first uniformly samples from one of the cases for which the green (or blue) marble would knock the gray marble into the gate, and then applies noise to the sampled velocity vector just like in Experiment 1. The model repeats this sampling procedure many times for the two candidate causes, and counts the proportion of cases for which the outcome in the simulation would have been different from what actually happened (see Eq. (1)). The model then uses this probability to predict both judgments of difficulty and judgments of blame. The more situations there are in which the blue/green marble knocks the gray marble into the gate, the less difficult the shot is predicted to be, and the more a player is predicted to be blamed for having failed to knock the marble into the gate.

Note that in the experiment, we asked participants whose shot was more difficult and which player was more to blame for the negative outcome. To fit participants' judgments, the model considers the probability that each marble would knock the gray marble into the gate, and then transforms these probabilities into a preference for one marble over the other via a soft-max decision function (Luce, 1959; Sutton & Barto, 1998). Because we fitted the temperature parameter in the soft-max function, this means that as long as the model's predictions are in the right direction, it will be able to capture participants' judgment. However, since the model assumes a direct mapping between difficulty and blame judgments, it is constrained to predict symmetrical responses to these questions.

To sum up, CSM makes the following predictions about participants' judgments in Experiment 3:

Hypothesis 3.1. Participants will judge that it will be more difficult for the player with the obstacle to hit the gray marble into the goal (i.e. the green player in Fig. 7).

Hypothesis 3.2. When both players failed to hit the gray marble into the goal, participants will judge that the player without the obstacle (i.e. the blue player in Fig. 7) is more to blame than the player with the obstacle.

Hypothesis 3.3. There will be a direct mapping between participants' difficulty and blame judgments. The easier they thought it was for one player compared to the other to hit the gray marble into the goal, the more that player should be blamed when the gray marble did not go in.

[12] There is no friction in our setting and the collisions are perfectly elastic, so the initial speed with which a marble moves does not matter.

[13] Fig. A2 in the appendix shows how the counterfactual probability is affected by setting a maximum on the number of wall collisions allowed. Intuitively, participants will not consider situations in which a colored marble knocks the gray one into the gate after a large number of collisions with the walls.

5.3. Results

Fig. 8 shows participants' difficulty and blame judgments. As predicted, participants indicated that it was more difficult for the player close to the obstacle to knock the gray marble through the gate (i.e. the green player in Fig. 7), $M = 86.28$, $SD = 21.84$ (Hypothesis 3.1). A directed one-sample t-test against the midpoint of the scale was significant, $t(103) = 16.94$, $p < .001$, $d = 1.66$. Conversely, participants blamed the player more who had a straight shot (i.e. the blue player in Fig. 7) $M = 29.04$, $SD = 20.75$ (Hypothesis 3.2). A directed one-sample t-test against the midpoint of the scale was significant, $t(103) = -10.30$, $p < .001$, $d = 1.01$.

There was no effect of question order on difficulty or blame judgments ($t(102) = 0.05$, $p = .961$ and $t(102) = -1.12$, $p = .267$, respectively). The faint lines in Fig. 8 indicate what pair of judgments each participant gave. 62 out of 104 participants indicated both that the shot is more difficult for the player close to the obstacle and that the player far from the obstacle is more to blame. 36 participants considered both players equally blameworthy, while only 8 participants considered both players' shots to be equally difficult (treating judgments within the range of 45 to 55 as indicating equality). Overall, this suggests that, unlike predicted in Hypothesis 3.3, there was no direct mapping between how participants judged difficulty and blame.

5.4. Discussion

The results of Experiment 3 show that when multiple potential causes failed to bring about an outcome, people blame the player more for whom it would have been easier to make the outcome happen. Participants judged that scoring is easier for the player who had a direct shot at the target compared to the player whose shot was obstructed by an obstacle. Correspondingly, when both players failed to score, participants were inclined to blame the player more who had an easier shot.

The CSM predicts this pattern by assuming that people mentally simulate what would have happened if each player had shot their marble differently. Given that the players' goal is to knock the gray marble into the goal, we assume that people are more likely to consider shots that involve few rather than many collisions with the walls prior to knocking the gray marble into the goal. Given this assumption, there is a greater chance that the marble without the obstacle will be successful than the marble with the obstacle.[14]

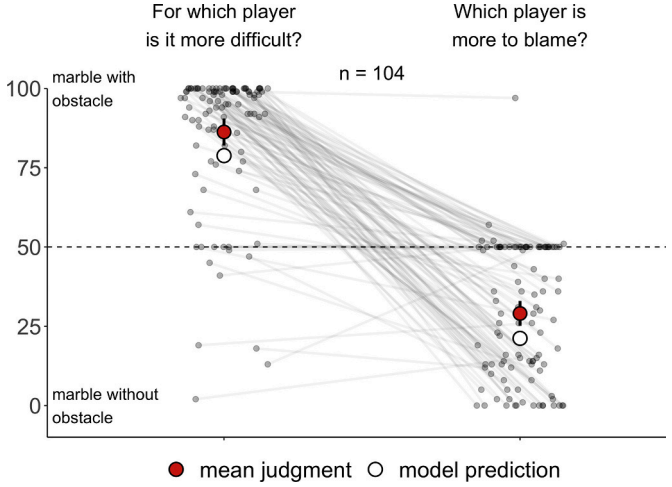

Fig. 8. Experiment 3 results: Participants' responses to the difficulty question (left) and the blame question (right). High ratings indicate a preference for the player whose marble was next to the obstacle (see Fig. 7), and low ratings a preference for the player without an obstacle. Large red points indicate mean judgments with 95% bootstrapped confidence intervals. Large white points indicate model predictions. Small black dots indicate individual responses (jittered along the x-axis for visibility). Lines connect individual difficulty and blame question responses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The CSM predicts a direct mapping from judgments of difficulty to judgments of blame, as it assumes that underlying each judgment is an assessment of the probability that a shot should have been successful. While this captures the main trends in the data (see the model predictions in Fig. 8), there was also an unanticipated asymmetry. While almost all participants agreed that scoring is more difficult for the player with the obstacle, a number of participants assigned equal blame to both players. Taking a look at the open-ended responses that participants provided at the end of the experiment revealed that some participants assigned blame in a purely outcome-based manner: Several participants stated explicitly that the shot was more difficult for one player, but that both players are still equally to blame since they both failed (e.g. "both players didn't knock the grey marble into the goal, equally to blame. but blue marble in harder position").[15] In contrast, many other participants explained their differential assignment of blame by reference to difficulty (e.g. "I thought the green played was more to blame because they had an easier path to the grey marble.").

Experiment 3 featured a single test clip. As such, it's not possible to rule out alternative hypotheses for how participants may have arrived at their judgments. Maybe participants based their judgments directly on features of the scene, such as the position of the barrier, or the number of times the balls collided, and did not consider counterfactual simulations. A challenge for such a feature-based account would be to explain why these features should matter for causal judgments. The CSM predicts that features like the position of the barrier matter because they affect how likely the outcome would have been different in relevant counterfactual simulations. Nevertheless, to rule out feature-based alternative accounts, future work needs to test the CSM in a wider variety of situations.

6. General discussion

In our everyday lives, we often cite things that *didn't happen* to explain things that *did happen*. For example, it seems perfectly fine to say that our beautiful orchids died because our neighbor did not keep his promise to water them while we were away, or that a pandemic led to a national disaster because the administration refrained from taking appropriate precautionary measures. How do people make causal judgments about things that did not happen?

In this paper, we developed an extension of the *counterfactual simulation model* (CSM) to explain people's causal judgments about omissions (Gerstenberg et al., 2021). The key idea is that people use their intuitive domain understanding to mentally simulate what would have happened if the omitted event had occurred. People believe that an outcome happened *because* of an omission the more certain they are that the outcome would not have happened if the omitted event had occurred. People's causal judgments are affected by their expectations about what will happen. According to the CSM, expectations modulate what counterfactuals people consider, which in turn affects their causal judgments. People are predicted to judge that an omission was more causal when the omitted event was expected to happen in a way that would have made a difference to the outcome. The results of three experiments showed that people's causal judgments about omissions are consistent with the CSM's predictions.

We broke the problem of explaining why something happened down into four sub-problems. First, the *variable selection problem* is about what

[14] Fig. A2 in the Appendix shows the number of successful shots for both marbles as a function of how many times each marble collides with one of the walls or the obstacle before knocking the gray marble into the goal.

[15] Participants' explanations may be seen in the analysis file posted online https://cicl-stanford.github.io/omission/.

variables are considered as candidate causes of the outcome (Halpern & Hitchcock, 2011; Woodward, 2015). The selection problem is particularly challenging when omissions are allowed to be causes. There are many factors that influence what variables come to mind (Branscombe, Owen, Garstka, & Coleman, 1996; Byrne, 2016; Mandel, 2003, 2011; N'gbala & Branscombe, 1995). In line with prior work, we believe that expectations about what normally happens guide people's selection of causes (Hesslow, 1988; Kahneman & Miller, 1986; Kahneman & Tversky, 1982; McGrath, 2005) as well as thoughts about what is to be learned from what happened, and what one could do to make a difference to the outcome in the future (Gerstenberg & Icard, 2019; Girotto et al., 1991; Hitchcock, 2012; Lombrozo, 2016; Phillips et al., 2019).

Second, the *specification problem* is about how the variables in the model carve up the event space. We focused on situations in which variables are binary and denote whether an event happened or did not happen. The specification problem is also particularly challenging when considering omissions. When an event happened, it is often straightforward to specify how that event might not have happened. However, when an event did not happen, it is less clear what the relevant contrastive event should look like (Halpern & Hitchcock, 2015; Schaffer, 2005, 2010). For example, when Tom shot Steve, it's easy to consider the event of Tom not shooting Steve. However, when Tom did not shoot Steve, it's less clear what one is supposed to imagine the alternative event of Tom shooting Steve (since there are many possible ways to shoot someone).

Third, the *evaluation problem* is about how to simulate the consequences of what would have happened if the event of interest had occurred (Gerstenberg & Tenenbaum, 2017; Kahneman & Tversky, 1982). A counterfactual model of causation like the CSM has to specify a mechanism that realizes the desired counterfactual, and then simulates what the outcome would have been. The CSM assumes that people use their intuitive understanding of physics to generate imagined interventions on the scene by imparting a force on a candidate causal object.

Finally, the *attribution problem* arises when there are several candidate causes and the question is to what extent each of the causes is responsible for bringing about the outcome (Gerstenberg & Icard, 2019; Gerstenberg & Lagnado, 2010; Icard et al., 2017; Lagnado et al., 2013). To address this problem, the CSM evaluates each of the candidate causes and then attributes causal responsibility based on how easy it is to imagine that a candidate cause could have made a difference to the outcome. The ease of bringing about the relevant counterfactual situation in which the outcome would have been changed is affected by expectations that bias the generation of counterfactuals.

Experiments 1 and 2 looked at how the specification problem and the evaluation problem affect judgments of omissive causation. In Experiment 1, the CSM addresses the specification problem by considering the possible ways in which Marble A could have collided with Marble B. How these counterfactuals are generated is affected by the expectations that participants have about what would happen. More specifically, the CSM models the influence of expectations by varying the degree of noise that is introduced into the sampling process of counterfactual simulations. The stronger the expectations, the less noise is introduced. As predicted by the CSM, Experiment 1 revealed an asymmetry: Marble A's not hitting Marble B was judged less causal when Marble B missed the gate compared to when Marble B went through the gate. Adding expectations increased both people's causal judgments as well as their subjective degree of belief that a collision would have changed the outcome. This effect was particularly strong when participants had expectations about social agents playing a marbles game. The CSM explains this effect of expectations by assuming that knowledge about intentions of agents limits what counterfactuals are considered. Our results thus add to previous research indicating that intentional actions signal higher causal stability compared to unintentional ones (Heider, 1958; Lombrozo, 2010), and that causal stability is indeed a relevant dimension that affects causal judgment (Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Lewis, 1986; Nagel & Stephan, 2016; Vasilyeva, Blanchard, & Lombrozo, 2018; Woodward, 2006).

One potential objection to our interpretation of the results from Experiment 1 is that the differences in causal judgments between the situations in which the marble went through the gate or missed the gate might be merely due to an inherent asymmetry between omissions that prevent and omissions that cause. Experiment 2 addressed this concern by contrasting causation and prevention in a situation in which the specification problem and evaluation problem did not arise. As predicted by the CSM, when it was clear that the outcome would have been different in the relevant counterfactual situation, participants' causal judgments were at ceiling for both causation and prevention by omission.

Experiment 3 focused on the attribution problem: how do people determine which out of several omissions was most responsible for the outcome? The results showed that people attributed more responsibility to an omission for which the counterfactual contrast that would have resulted in a positive outcome was easier to imagine. Specifically, when both players missed their shot in a cooperative two-player marble game, the player who had the easier shot was blamed more for the negative team outcome. This intuitive result highlights the importance of how the relevant counterfactual contrast is specified. In the actual situation, both players missed their shot. One way to specify the counterfactual contrast would have been to say, if the player's marble had hit the target marble, it would have gone through the gate. This specification, however, would not predict a difference between the two players. The target marble is positioned right in front of the goal such that it's almost certain to go into the goal if it was struck. Specifying the counterfactual as "if the target marble had been hit" would not predict any difference between the two players regardless of whether their shot was easy or difficult. Instead, participants' judgments are consistent with the counterfactual contrast being specified as "if the player had shot the marble differently" (or "if another – reasonably skilled – player had shot the marble"). This way, the model can account for the observed results. The probability that the outcome would have been positive had the player shot the marble differently is greater for the player with the easy shot compared to the player with the difficult shot. Future research needs to investigate what determines how people construe the relevant counterfactual contrasts.

Our experiments did not address the selection problem. In all of our experiments, we explicitly asked participants about the candidate causes. What triggers the search for causal explanations (Hastie, 1984; Kanazawa, 1992; Weiner, 1985; Wong & Weiner, 1981), how people naturally and spontaneously construct the causal models that support such explanations (Hagmayer & Osman, 2012; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Sloman, 2005), and how this model-building process is guided by what they know and what their goals are (Gerstenberg & Icard, 2019; Hilton, 1990) remain important questions for future research.

In the remainder, we will discuss how the force dynamics model (Wolff et al., 2010) and the mental model theory (Khemlani et al., 2018) might account for the results of our experiments, what the CSM has to say about the actual process by which people are making causal judgments about omissions, how our model could be extended to capture different causal expressions about what happened, and what our model has to say about omissive causation outside of the physical setting.

6.1. Alternative accounts

The *force dynamics model* (Wolff et al., 2010) predicts that omissive causation is grounded in the removal of an actual or anticipated force. In Experiments 1 and 3, no actual forces are exchanged since the candidate marbles either do not move at all (Experiment 1) or they miss the target marble (Experiment 3). In Experiment 2, the wall blocks the marble from going through the gate in one of the two clips. In this case, there is an actual force that prevents the outcome from happening. To account for

the rest of the clips, however, the force dynamics model would have to rely on the notion of an anticipated force that was not realized. The notion of an unrealized anticipated force is of course a counterfactual one. In that sense, the CSM and the force dynamics model converge here. However, the two accounts still differ with respect to how people's domain knowledge is represented. The force dynamics model uses vector algebra to represent people's knowledge, whereas the CSM assumes that people's physical knowledge resembles in important ways the workings of a physics engine of the kind that are used to model realistic physical interactions in computer games (Gerstenberg & Tenenbaum, 2017; Goodman, Tenenbaum, & Gerstenberg, 2015; Ullman et al., 2017; but see also Ludwin-Peery, Bramley, Davis, & Gureckis, 2021).

In the experiments reported in Wolff et al. (2010), the physical interactions play out in 1D, and the specification problem is fairly minimal – the relevant counterfactual contrast is the absence of a candidate cause (i.e. what would have happened if one of the agents had been removed). The evaluation problem is also fairly minimal, although participants might be somewhat uncertain about the magnitude of the actual or anticipated force vectors which can make a difference to what would have happened. It is unclear how the force dynamics model would account for the results of our Experiment 1. For the CSM, the effect that the clip has on participants' judgments follows directly from the evaluation problem: it's more likely that a marble that's on target would have missed the gate if it had been hit, than it is for a marble that originally was not on target to have gone through the gate had it been hit. The CSM suggests a concrete mechanism for how expectations affect the consideration of counterfactuals that the force dynamics model lacks at this point. Incorporating expectations is also critical for predicting that one player is blamed more than the other for missing the target in Experiment 3.

The *mental model theory* (Khemlani et al., 2018) predicts that omissive causation is grounded in the representation of concrete possibilities. "Causing" by omission versus "enabling" by omission are defined in terms of a set of logical possibilities that are consistent with each concept. Since the model requires the existence of multiple possibilities to apply, it is better suited to making predictions about general causal relationships rather than particular causal events. In our experiments, participants only viewed a single video clip rather than being exposed to multiple different clips depicting different possibilities which the mental model theory requires for its predictions.

Neither the force dynamics model nor the mental model theory make quantitative predictions. Quantitative predictions are important, however, as they allow for more rigorous tests of the theory. The CSM makes quantitative predictions and these predictions have been shown to accurately capture participants' judgments about commissive causes (Gerstenberg et al., 2017, 2021). Here, we have shown how participants' judgments about omissive causes are also consistent with the CSM. Our experiments were limited in that they only featured a small number of situations which precluded a more quantitative analysis of the model predictions. For a more stringent test of the model, future work needs to manipulate in a more continuous fashion the physical settings of the scene as well as participants' expectations about what will happen.

6.2. The process of making causal judgments about omissions

The CSM proposes that people make causal judgments about omissions by comparing what actually happened with a mental simulation of what would have happened in a counterfactual situation in which the event of interest had taken place. Gerstenberg et al. (2017) demonstrated that when participants are asked to make causal judgments about commissions, they spontaneously engage in counterfactual simulation as evidenced by their eye-movements. Participants do not just look at what actually happened, they try to mentally simulate what would have happened if the causal event of interest had not taken place.

Here, we build on this work to suggest a model for how people make causal judgments about omissions, and for how expectations shape the way in which people consider counterfactual possibilities. In order to make these ideas concrete, we had to make a number of implementation decisions. Our model generates counterfactual possibilities by first sampling a situation in which the counterfactual event of interest would have happened, and then somewhat perturbing that sample with noise. The model assumes that prior expectations affect how much noise is applied to these generated samples (with higher expectations that a causal event would have happened resulting in less noise in the counterfactual simulations). This is of course not the only way in which the idea that expectations affect the generation of counterfactuals could be implemented, and we are not strongly committed to the particular implementation that we chose. What we are committed to is the more general idea that causal judgments about omissions can be understood by considering counterfactuals on the generative model of the situation, and that expectations affect what counterfactuals are generated. Eye-tracking may help to gain more direct insights into what counterfactuals people are considering (see Gerstenberg et al., 2017).

6.3. The language of omissions

In our experiments, we asked participants to what extent they agreed with statements that the outcome happened *because* of an omission. We did not ask participants whether the omission *caused* the outcome. This was a deliberate choice. For example, in Experiment 1, the prompt was "Marble B went through the gate, because Marble A did not hit Marble B." We could have also asked participants to judge whether Marble A's not hitting Marble B caused Marble B to go through the gate. Our intuition is that while the "because" variant sounds fine and natural, the "caused" variant does not. We are not the first to notice this. In fact, Beebee (2004) argues that there is a fundamental difference between causation (for which "caused" is appropriate) and causal explanations (for which "because" is appropriate). In a series of experiments, Livengood and Machery (2007) find that participants often agree more with "because" variants compared to "caused" variants in cases of omissive causation.

For the purposes of this paper, we laid out merely the core of the counterfactual simulation model. However, the CSM has also been applied to capturing participants' judgments in scenarios involving multiple causes (Gerstenberg et al., 2021). What these scenarios revealed is that people's causal judgments are sensitive to multiple aspects of causation. People care not only about *whether* an outcome happened but also about *how* it came about. Traditional counterfactual theories of causation focus on whether-causation. They ask the question: Did the presence versus absence of the cause make a difference to whether or not the outcome happened? However, these theories have difficulty accounting for a number of cases in which multiple causes are causally relevant for the outcome. For example, consider a simple causal chain in which Marble A knocks into Marble B which subsequently knocks Marble C into the gate. Intuitively, Marble B was causally relevant for Marble C's going through the gate. But a simple counterfactual test fails here: Marble C would have gone through the gate even if Marble B had not been present in the scene (because Marble A would have knocked Marble C into the gate in this case).

The CSM accounts for cases like these by postulating another aspect of causation: how-causation. How-causation is revealed through a different counterfactual test. Rather than considering what would have happened if the cause had been present versus absent, for how-causation the CSM considers what would have happened if the candidate cause had been subtly changed. Concretely, one might imagine this change as

a small perturbation to the marble's position. A candidate cause is a how-cause if it's the case that the outcome would have been different had the cause been changed.[16] So, in the causal chain, Marble B is not a whether-cause of Marble C's going through the gate but it is a how-cause (whereas Marble A is both a how-cause and a whether-cause).

We believe that considering these different aspects of causation may help to explain why "because" but not "caused" statements are appropriate for omissions. Omissions can be whether-causes. As shown in our experiments, the absence versus the presence of the cause makes a difference to whether or not the outcome happened. However, omissions cannot be how-causes. The idea of considering how an omission could have been slightly different from how it actually happened is arguably ill-defined. Non-events do not have a spatio-temporal signature. They do not happen in space or time (Bernstein, 2015; Lewis, 2000, 2004), and so we cannot consider how these non-events could have happened slightly differently from how they actually did (not).

The difference between the "because" and "caused" statements might thus lie in what requirements the candidate cause needs to fulfill. For "because" statements, it seems sufficient that the candidate cause was a whether-cause of the outcome, whereas for "caused" statements the candidate should be both a whether-cause *and* a how-cause of the outcome. In fact, Beller, Bennett, and Gerstenberg (2020) have shown in recent work that the different aspects of causation help explain differences between causal expressions such as "caused", "enabled", "affected" and "made no difference". In this work, participants watched video clips of physical interactions similar to the ones we used here, and were asked to choose which expression best describes what happened. Beller et al. propose a literal semantics of the different causal expressions using the CSM's aspects of causation and show that for participants' interpretation of "caused" both how-causation and whether-causation is critical, whereas for "enabled" whether-causation matters, and for "affected" how-causation is important. In future work, we will investigate what causal expressions people use to describe situations of omission (cf. Khemlani et al., 2018).

6.4. Omissions beyond the physical

In this paper, we applied the counterfactual simulation model to predict participants' judgments about omissions in relatively simple physical settings. By restricting ourselves to this well-defined setting, we were able to postulate a concrete mechanism of how people may generate counterfactual simulations, and of how prior expectations influence the consideration of counterfactuals.

However, people regularly make causal judgments in situations that are much more complex, with limited information, and where it would be impossible to mentally simulate exactly how certain counterfactuals might play out. We mentioned the example of a pandemic leading to a national disaster because the administration refrained from taking appropriate precautionary measures. Our intuition is that people can still make causal judgments in complex situations like these because they have the ability to construct mental models that abstract away many of the low-level details of the situation (see Beckers & Halpern, 2019; Ullman et al., 2017). So instead of mentally simulating what actions each person would have taken, one would need to abstract away from this low level, and then consider the counterfactual dependence between the variables of interest on a higher level of abstraction.

As our introductory example demonstrates, omissions are particularly relevant in human interaction, especially in morally or legally charged situations where we have clear expectations about what a person should have done. For example, in bystander situations we hold people morally responsible for not acting when they should have (Darley & Latané, 1968; Fischer et al., 2011). What, if anything, can the model we propose here say about people's causal judgments outside of the physical domain?

We believe in the power of the general idea that causal judgments are well-understood in terms of counterfactual operations defined over generative models of the domain (Gerstenberg et al., 2021; Gerstenberg & Tenenbaum, 2017). People's domain understanding will dictate both what counterfactuals come to mind, as well as how to simulate what would have happened had things played out differently. As we have seen in Experiment 3, evaluating agents' actions requires different counterfactuals from considering merely physical events. Participants' blame judgments where consistent with the assumption that they are imagining what a reasonable person would have done in the same situation. The person for whom it's easier to imagine that they could have succeeded is blamed more. Relatedly, Jara-Ettinger et al. (2015) have shown that even toddlers already evaluate an agent who refused to help more negatively when helping would have been easy.

The reasonable person test is a common procedure in the law to evaluate legal responsibility in cases of negligence, in which harm resulted from a person's failure to act (Gerstenberg et al., 2018; Green, 1967). Much work in moral psychology (Clarke, 1994; DeScioli, Bruening, & Kurzban, 2011; Royzman & Baron, 2002; Waldmann, Nagel, & Wiegmann, 2012) and decision-making (Anderson, 2003; Baron & Ritov, 1994, 2004; Byrne, 2005, 2016; Ritov & Baron, 1992; Spranca, Minsk, & Baron, 1991; Zeelenberg, Van den Bos, Van Dijk, & Pieters, 2002) has shown that people make different judgments about and draw different inferences from omissions versus commissions. For example, Greene et al. (2009) showed that people were less likely to consider a person's action morally acceptable when it directly impacted the victim (e.g. via pushing) compared to when there was no physical contact (cf. De Freitas & Alvarez, 2018; Iliev, Sachdeva, & Medin, 2012; Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021). This finding is consistent with the idea that, all else being equal, actions strike us as more causal than non-actions because for actions there is less uncertainty about what the relevant contrast is. More research is required into the factors that influence what counterfactuals come to mind, and how these counterfactuals in turn may influence causal judgments, moral evaluations, and decisions. By developing computational models of how people make decisions and take actions, and by combining these models with general tools for doing counterfactual inference (e.g. Evans, Stuhlmüller, Salvatier, & Filan, 2017; Perov et al., 2020; Tavares, Koppel, Zhang, & Solar-Lezama, 2019), we will be able to build and test models that make quantitative predictions about causal judgments outside of the physical domain we considered here.

7. Conclusion

Omissions have a complicated causal status. How can something that did not happen be the cause of something else happening? Much prior work has argued for the role of mental simulation in causal judgments (e.g. Goldvarg & Johnson-Laird, 2001; Kahneman & Tversky, 1982), and for the idea that people's beliefs about what is normal affect their causal judgments via shaping what counterfactuals come to mind (e.g. Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Kominsky & Phillips, 2019). In this paper, we have presented the first concrete implementation of these ideas for handling causal judgments about omissions in a physical domain. The counterfactual simulation model (CSM) predicts that people compare what actually happened with what would have happened in a simulated counterfactual simulation (see also Gerstenberg et al., 2021), and that prior expectations influence how these counterfactual simulations are generated.

[16] Note that for how-causation the outcome event is construed finely as a continuous variable that includes information about when and where the event happened, whereas for whether-causation the outcome event is construed coarsely as a binary variable that merely represents whether or not the outcome happened (e.g. whether the marble went through the gate or missed it).

Acknowledgments

We thank Pascale Willemsen for her contributions in the early stages of this project, Shardul Chiplunkar for help with implementing the model, and Henrik Singmann for advice with statistical analyses. We thank Bryce Linford, Shardul Chiplunkar, and Xi Jia Zhou for providing feedback on the manuscript. We thank Sunny Khemlani and one anonymous reviewer for the constructive feedback that helped to improve the paper. Part of this work has appeared in the *Proceedings of the Cognitive Science Conference*: Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in Inaction: Counterfactual Simulation and Causation by Omission. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, Austin, TX, 2017 (pp. 1132–1137). Cognitive Science Society.

This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216, and by the Leibniz Association through funding for the Leibniz ScienceCampus Primate Cognition.

Appendix A

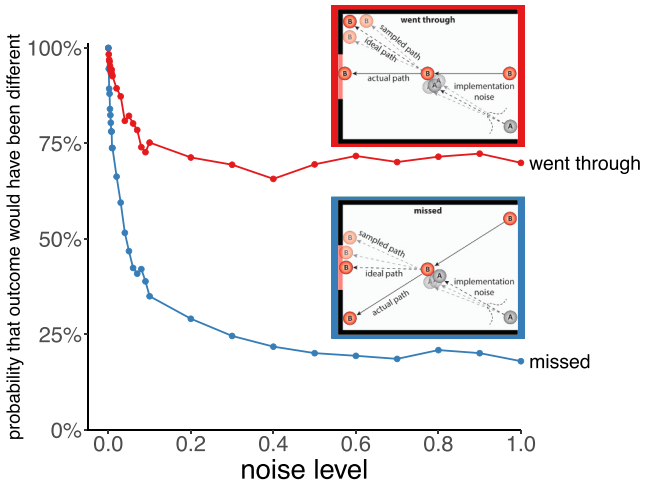

Fig. A1. Experiment 1 model results: Each point indicates the probability that the outcome would have been different from what actually happened for different amount of noise applied to marble A's initial trajectory. The blue points are for the clip in which marble B actually missed the gate, and the red points are for the clip in which marble B actually went through the gate. Each point is based on 1000 simulated model runs. For each run, the model first uniformly samples a case in which marble A was shot in a way such that the counterfactual outcome was different from what actually happened (e.g. such that marble B would have gone through the gate when it actually missed). The model then applies noise to that velocity vector, and records the outcome of the simulation. When the noise is very small, the outcome would almost always be different from what actually happened no matter whether marble B originally missed or went through the gate. However, as the noise increases, the probabilities of the relevant counterfactual for the two situations separate. The probability that the counterfactual outcome would have been different from what actually happened decreases more rapidly for "missed" compared to "went through' 'as the noise level increases. This indicates that making marble B miss (when it originally went through) is more robust to noise, than making marble B go through (when it originally missed). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

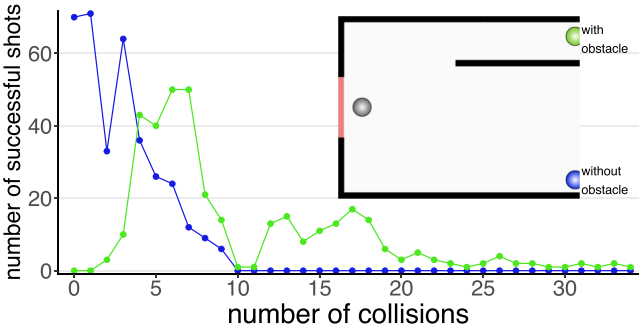

Fig. A2. Number of successful shots for the marble without the obstacle (blue), and the marble with the obstacle (green), as a function of how many collisions happened before the marble knocked the gray marble into the gate (*x*-axis). For the marble without the obstacle, there are many more successful shots for a low number of collisions. For the marble with the obstacle, it needs at least two collisions with the walls before it can knock the gray marble into the gate. While there are in fact an almost equal number of shots that would be successful for both the marble with and without the obstacle overall, it's plausible to assume that participants have a tendency to simulate paths with lower number of collisions for which the marble without the obstacle is more likely to generate successful outcomes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Achinstein, P. (1983). *The nature of explanation*. Oxford University Press.

Anderson, C. J. (2003). The psychology of doing nothing: Forms of decision avoidance result from reason and emotion. *Psychological Bulletin, 129*(1), 139–167.

Baron, J., & Ritov, I. (1994). Reference points and omission bias. *Organizational Behavior and Human Decision Processes, 59*(3), 475–498.

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*(2), 74–85.

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 2678–2685).

Beebee, H. (2004). Causing and nothingness. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 291–308). MA: MIT Press Cambridge.

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In *Proceedings of the 42nd annual conference of the cognitive science society*.

Bello, P., & Khemlani, S. S. (2015). A model-based theory of omissive causation. In R. Dale, et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

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

Bernstein, S. (2015). The metaphysics of omissions. *Philosophy Compass, 10*(3), 208–218.

Bohner, G., Bless, H., Schwarz, N., & Strack, F. (1988). What triggers causal attributions?. The impact of valence and subjective probability. *European Journal of Social Psychology, 18*(4), 335–345.

Branscombe, N. R., Owen, S., Garstka, T. A., & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome?.1. *Journal of Applied Social Psychology, 26*(12), 1042–1067.

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

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

Clarke, R. (1994). Ability and responsibility for omissions. *Philosophical Studies, 73*(2), 195–208.

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). Preemptive prevention. *The Journal of Philosophy, 97*(4), 223.

Danks, D. (2017). Singular causation. In M. Waldmannn (Ed.), *The oxford handbook of causal reasoning* (pp. 201–215). Oxford University Press.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology, 8*(4), 377–383.

De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition, 178*, 133–146.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review, 127*(July (4)), 591–621.

DeScioli, P., Bruening, R., & Kurzban, R. (2011). The omission effect in moral cognition: Toward a functional explanation. *Evolution and Human Behavior, 32*(3), 204–215.

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

Dowe, P. (2001). A counterfactual theory of prevention and "causation" by omission. *Australasian Journal of Philosophy, 79*(2), 216–226.

Evans, O., Stuhlmüller, A., Salvatier, J., & Filan, D. (2017). *Modeling agents with probabilistic programs*. http://agentmodels.org Accessed 22.05.17.

Fair, D. (1979). Causation and the flow of energy. *Erkenntnis, 14*(3), 219–250.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using bayesian networks. *Cognitive Science, 37*(1), 61–102.

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., et al. (2011). The bystander-effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*(4), 517–537.

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., Ejova, A., & Lagnado, D. A. (2011). Blame the skilled. In C. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 720–725). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgment. *Psychological Review*.

Gerstenberg, T., & Icard, T. F. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*.

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

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

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

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

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

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

Green, D. W. (2008). Persuasion and the contexts of dissuasion: Causal models and informal arguments. *Thinking & Reasoning, 14*(February (1)), 28–59.

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). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology, 11*, 1069.

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

Hagmayer, Y., & Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General, 138*(1), 22–38.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals*. MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies, 132*, 109–136.

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

Halpern, J. Y., & Hitchcock, C. (2011). Actual causation and the art of modeling. In R. Dechter, H. Geffner, & J. Y. Halpern (Eds.), *Heuristics, probability and causality: A tribute to Judea Pearl* (pp. 316–328). College Publications.

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.

Harinen, T. (2017). Normal causes for normal effects: Reinvigorating the correspondence hypothesis about judgments of actual causation. *Erkenntnis*.

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. New York: Oxford University Press.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Personality and Social Psychology, 46*(1), 44–56.

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

Henne, P., Bello, P., Khemlani, S. S., & Brigard, F. D. (2019). Norms and the meaning of omissive enabling conditions. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the cognitive science society*. Montreal, Canada.

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

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy, 95*(2), 270–283.

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.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin, 107*(1), 65–81.

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

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. (1995). The mishap at reichenbach fall: Singular vs. general causation. *Philosophical Studies, 78*(June (3)), 257–291.

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

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.

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.

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

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kanazawa, S. (1992). Outcome or expectancy?. Antecedent of spontaneous causal attribution. *Personality and Social Psychology Bulletin, 18*(6), 659–668.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*(1), 227–254.

Khemlani, S., Bello, P., Briggs, G., & Harner, H. (2020). Much ado about nothing: The mental representation of omissive relations. *PsyArXiv*.

Khemlani, S., Wasylyshyn, C., Briggs, G., & Bello, P. (2018). Mental models and omissive causation. *Memory & Cognition, 46*(8), 1344–1359.

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

Kirfel, L., Icard, T. F., & Gerstenberg, T. (2020). Inference from explanation. *PsyArXiv*.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science, 43*(11), e12792.

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

Kunda, Z. (1987). Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology, 53*(4), 636–647.

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.

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

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

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, D. (2004). *Void and object*.
Livengood, J. (2011). Actual causation and simple voting scenarios. *Noûs*, 1–33.
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.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*(4), 303–332.
Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences, 20*(10), 748–759.
Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review, 122*(4), 700–734.
Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley.
Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology, 127*, 101396.
Mandel, D. R. (2003). Judgment dissociation theory: An analysis of differences in causal, counterfactual and covariational reasoning. *Journal of Experimental Psychology: General, 132*(3), 419–434.
Mandel, D. R. (2011). Mental simulation and the nexus of causal and counterfactual explanation. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology* (pp. 146–170). Oxford University Press.
McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of Personality and Social Psychology, 79*(5), 677–689.
McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies, 123*(1), 125–148.
Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science, 71*(5), 820–832.
Menzies, P. (2006). *A structural equations account of negative causation*.
Mercier, H., & Sperber, D. (2011). Why do humans reason?. Arguments for an argumentative theory. *Behavioral and Brain Sciences, 34*(March (2)), 57–74.
Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 806–812).
N'gbala, A., & Branscombe, N. R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology, 31*(2), 139–162.
Nyberg, E. P., Nicholson, A. E., Korb, K. B., Wybrow, M., Zukerman, I., Mascaro, S., et al. (2021). BARD: A structured technique for group elicitation of bayesian networks to support analytic reasoning. *Risk Analysis*, (June).
Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.
Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.
Perov, Y., Graham, L., Gourgoulias, K., Richens, J., Lee, C., Baker, A., et al. (2020). Multiverse: causal reasoning using importance sampling in probabilistic programming. In *Symposium on advances in approximate bayesian inference* (pp. 1–36).
Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of Personality and Social Psychology, 100*(1), 30–46.
Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences, 23*(12), 1026–1040.
Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science, 83*(December (5)), 721–732.
Ritov, I., & Baron, J. (1992). Status-quo and omission biases. *Journal of Risk and Uncertainty, 5*(1).
Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research, 15*(2), 165–184.
Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science, 61*(2), 297–312.
Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition, 156*, 164–176.
Sanna, L. J., & Turley, K. J. (1996). Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. *Personality and Social Psychology Bulletin, 22*(9), 906–919.
Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science, 67*(2), 285.
Schaffer, J. (2005). Contrastive causation. *The Philosophical Review, 114*(3), 327–358.
Schaffer, J. (2010). Contrastive causation in the law. *Legal Theory, 16*(04), 259–297.
Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. USA: Oxford University Press.
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.
Sosa, F. A., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *PsyArXiv*.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology, 27*(1), 76–105.
Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation: A computational model. *Cognitive Science, 44*(7), e12871.
Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science, 10*(1), 242–257.
Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge University Press.
Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*(1), 49–100.
Tavares, Z., Koppel, J., Zhang, X., & Solar-Lezama, A. (2019). *A language for counterfactual generative models*.
Turnbull, W., & Slugoski, B. R. (1988). Conversational and linguistic processes in causal attribution. In D. J. Hilton (Ed.), *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 66–93). Brighton, UK: Harvester Press.
Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences, 21*(9), 649–665.
Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science, 42*(4), 1265–1296.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.
Weiner, B. (1985). "spontaneous" causal thinking. *Psychological Bulletin, 97*(1), 74–84.
Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*(2), 161–169.
Willemsen, P. (2018). Omissions and expectations: A new approach to the things we failed to do. *Synthese, 195*(4), 1587–1614.
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., Hausknecht, M., & Holmes, K. (2011). Absent causes, present effects: How omissions cause events. In J. Bohnemeyer, & E. Pederson (Eds.), *Event representation in language and cognition*. Cambridge, UK: Cambridge University Press.
Wong, P. T., & Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *Journal of Personality and Social Psychology, 40*(4), 650–663.
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. (2015). The problem of variable choice. *Synthese, 193*(4), 1047–1072.
Zeelenberg, M., Van den Bos, K., Van Dijk, E., & Pieters, R. (2002). The inaction effect in the psychology of regret. *Journal of Personality and Social Psychology, 82*(3), 314–327.
Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition, 125*(3), 429–440.