Judgments of actual causation approximate the effectiveness of interventions

Adam Morris* Jonathan Phillips Thomas Icard
Harvard University Harvard University Stanford University
Joshua Knobe Tobias Gerstenberg Fiery Cushman
Yale University Massachusetts Institute of Technology Harvard University

Abstract

When many things contribute to an outcome, people consistently judge certain ones to be the outcome's "actual" cause. For instance, people believe the lit match, not the surrounding oxygen, was the cause of the fire. Why? Here, we offer a functional account of actual causation: Repeatedly judging whether something (e.g. the match) was the actual cause of an outcome (e.g. the fire) helps compute the probability that introducing it would produce the outcome. In other words, judgments of actual causation accumulate evidence about the effectiveness of potential interventions. We offer a formal account of this process, and show how it explains three basic qualitative features of causal judgment: why actual causes tend (1) to be necessary, (2) to be abnormal (the "abnormal selection" effect), and (3) to lack abnormal counterparts (the "supersession" effect). We show that this approach — which we call the "Sample-based Approximation Method for Predicting the Likelihood of Effectiveness", or the SAMPLE approach — makes quantitative predictions that closely match participants' judgments in a novel experiment.

I. Introduction

If an ancient forest burned down, we would naturally ask: "What caused the fire?" And in certain simple cases, we would all agree on the answer. If a careless camper dropped a match on dry leaves, we would judge her the cause. If the leaves were struck by lightning, we would point to the lightning strike.

But, on reflection, a multitude of factors contributed to the fire. The lack of rain; the National Park Service's decision to allow campers in this forest; the store owner selling the match to the camper; the oxygen in the air—all these were critical antecedents to the blaze, and yet people would not typically consider any of them to be the "actual" cause of the event. (Imagine a news anchor proclaiming that a forest fire was caused by oxygen in the air; it seems absurd.) Dry weather and oxygen recede in our minds as background conditions, while campers and lightning stand out as "actual" causes of the fire.

This phenomenon is widespread: When multiple variables contribute to an outcome, humans judge certain variables to be the actual cause(s) of the outcome (Hart & Honoré, 1959/1985; Hitchcock & Knobe, 2009; Mackie, 1974).[1] They believe the fire was caused by the careless camper; the political victory by the swing state; the car accident by the fog; and so on. Judgments of actual causation pervade human thought and discourse.

But while causal judgments are ubiquitous, their psychological basis remains mysterious. Why do people consistently foreground certain variables as actual causes, and relegate others to the status of mere background conditions

*All correspondence should be addressed to adam.mtc.morris@gmail.com.

[1] What we call actual causation is sometimes called "specific" causation (Sloman, Barbey, & Hotaling, 2009; Walsh & Sloman, 2011), "singular" causation (Hitchcock, 2017), or "causal selection" (Hitchcock, 2007). Also, we sometimes describe actual causation as applying to variables (e.g. `lit match` and `oxygen`), but other times we describe it as applying to events. We treat these as equivalent; for our purposes, the variable `lit match` is a cause just in case the event of someone lighting a match is a cause.

(Cheng & Novick, 1991; Hesslow, 1988; Hilton, 1990)? Why not treat all contributing variables equally?

In this paper, we offer a functional account of certain key aspects of actual causation judgments. We propose that they implement an approximation which assists in future decision-making by identifying effective ways to produce an outcome. Concretely, judging whether a match was the actual cause of a fire across various situations helps approximate the probability that lighting a match, in future situations, would produce a fire. We derive an algorithm that accomplishes this goal, accumulating evidence from past observations in order to guide future planning. Then, we show how some basic features of this algorithm can capture important properties of causal judgment that are already well-documented. We also provide new experimental evidence testing some of its unique predictions. Taken together, our findings suggest that human causal judgment is designed, in part, to accumulate evidence across diverse cases in order to approximate the general effectiveness of future interventions.

Actual causation as a sampling approximation Our model is inspired by recent philosophical treatments of causal judgment. These argue that people select actual causes that are "effective interventions"—i.e., good variables to change if you want to produce an outcome (Hitchcock, 2012; Hitchcock & Knobe, 2009; Woodward, 2003). Consider an admissions officer at a college. When she admits a student, she might judge that certain qualities—like an engaging essay—were the actual cause of the student's admission. (In contrast, she would probably not judge that filing the application by midnight on Dec. 31 was the actual cause of admission.) Moreover, the officer believes that certain interventions—like hiring a writing coach—are effective methods to get her own children into college. The crux of these recent proposals is that these kinds of judgments are related: The admissions officer judges the engaging essay causal precisely because making your essay engaging is generally a good way to make sure you get into college. Put simply: intervening to produce an outcome, and judging what caused the outcome, are intimately related.

We build upon this idea, but also diverge from recent approaches in two key ways.

First, we argue for a specific ordering of these judgments—a flow of information from one to the other: Knowledge of the optimal interventions does not guide causal judgment; rather, causal judgments accumulate knowledge about the optimal intervention. In other words, the admissions officer doesn't believe that an engaging essay caused the applicant's acceptance because she has identified good writing as an effective way to get into college. Rather, she identifies that good writing is an effective way to get into college because she has often judged it to be causal in the past.

The primary contribution of this paper is to offer a formal model of how and why this works—both the algorithm we use to identify variables like "good writing" as the cause of a specific applicant's success, and the reason that this algorithm helps us to identify the general strategies for successful interventions in the future. More specifically, we show that if actual causal judgments take a certain form, then averaging them across past episodes approximates the effectiveness of potential interventions. This form turns out to capture three salient features of actual causal judgments: their focus on variables which were necessary to produce the outcome, abnormal variables, and variables with normal counterparts. Our model thus offers a functional justification for these well-known features of actual causal judgments.

This perspective immediately raises a more general question: Of all the ways of planning to get someone into college, why rely on accumulated causal judgments of *past* applicants? In other words, of all the ways of identifying optimal interventions, why employ causal judgments? Even if such an approach could work, it is not obvious why it would be so ubiquitous, or even especially favored. There are certainly alternatives—for instance, one could rely ex-

clusively on a model of the *current* college application season to project the likely outcomes of different strategies.

Our second contribution in this paper is to propose an answer to this question: computational efficiency. In any causal system of real-world complexity, computing the effectiveness of an intervention is difficult—often prohibitively so (cf. Meder, Gerstenberg, Hagmayer, & Waldmann, 2010). A parent has an unbounded multitude of ways to help her child get into college, and exhaustive search over every possible strategy is not feasible. She must find a way to orient her evaluation toward strategies that are likely to be useful—strategies such as investing heavily in writing improvement. But this requires her to have a preexisting representation of what kinds of strategies work "in general". More abstractly, tractable planning depends on a representation of what interventions tend to produce desirable effects in general—i.e., averaging over a range of plausible circumstances (Hitchcock, 2012).

We propose that actual causal judgments make this computation tractable. A common solution to difficult computations is to employ a sampling approximation. Instead of explicitly enumerating over all the possible states of the system at decision time, a person can rely on her past experiences with the system—"samples" of that system, generated by the world and experienced by the person. By analogy, humans and non-humans sometimes solve decision-making problems not by undertaking the computational effort of computing expected values from their representation of a complex system, but by approximating value from their history of reward (Dolan & Dayan, 2013; Schultz, 2002). This works because the expected value of a future action can be estimated as the average of its value in past episodes.

We apply a similar idea to causal judgment. For instance, the admissions officer can consider the specifics of many past applicants, each of which represents a sample applicant situation. By analyzing the causal role of a variable (e.g. good writing) in these episodes as she experiences them, the person can later obtain an estimate of the "average", general role of a variable in the causal system. This, in turn, constitutes a useful starting point as she plans an admissions strategy for her own children.

In sum, we propose that judgments of actual causation help implement a sampling approximation to the overall effectiveness of an intervention. By repeatedly judging whether C was the actual cause of an outcome in samples of a causal system, people can obtain an estimate of the average likelihood that intervening on C would produce the outcome.

This paper proceeds as follows. We first review three prominent empirical patterns in actual causal judgments, and summarize previous approaches to understanding them. We next present our interventionist approach, and show how, in a simple but common decision problem, actual causation judgments can help approximate the effectiveness of potential interventions by adopting these three patterns. Above and beyond simply accounting for these patterns, our approach provides a functional justification for them. We then present a behavioral experiment testing a key component of our model and compare our model's performance to other extant models of actual causation. Finally, in the discussion, we consider how the principles of our approach may generalize to more complex decision problems.

II. The problem of actual causation

There are at least two questions to answer about the psychology of actual causation. First, *which* variables do people select as actual causes? What distinguishes the camper's dropped match from other variables that seem to be irrelevant (e.g. the color of the camper's shirt), or mere background conditions (e.g. the oxygen in the air)? This matter is well studied. Below, we review three prominent empirical patterns in actual causal judgments: "necessity", "abnormal selection", and "supersession".

Second, *why* do people select certain causes as actual (Hitchcock & Knobe, 2009; Woodward, 2014)? If you know that both a match and oxygen contribute to producing a fire, what is gained by differentiating between the actual cause—the match—and background conditions like oxygen? This question is less commonly addressed, and will motivate our approach.[2]

i. The *which* problem: three patterns in actual causal judgments

Although ordinary people's judgments of actual causation are complex, several clear patterns emerge.

Necessity The first pattern is that actual causes tend to be things that were *necessary* to produce the effect. Put simply, if the cause hadn't happened in the relevant situation, the effect wouldn't have either. For instance, if the camper hadn't dropped a match, the forest would not have burned down; if the student hadn't written an engaging essay, she would not have been accepted; and so on. The fact that the camper's match and the student's essay were necessary for their outcomes seems to be a key determinant of their causal status (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014; Halpern & Pearl, 2005; Hart & Honoré, 1959/1985; Kahneman & Tversky, 1982; Lewis, 1973; Mackie, 1974; Wells & Gavanski, 1989; Woodward, 2003). In contrast, the camper's decision to wear a blue shirt was not necessary for the fire, and so it is not considered a cause.

Necessity plays a central role in theories of causation across diverse fields, from philosophy (Mackie, 1974) to law (Hart & Honoré, 1959/1985) and computer science (Chockler, Halpern, & Kupferman, 2008). Its role in causal judgment is also supported by a wealth of psychological evidence (Gerstenberg et al., 2012; Gerstenberg, Goodman, et al., 2015; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017; Kahneman & Tversky, 1982; Wells & Gavanski, 1989). In sum, there is widespread agreement that necessity, in some form, is important for judgments of actual causation (Icard, Kominsky, & Knobe, 2017).

Abnormal selection The second pattern is that we tend to select *abnormal* events as the actual causes of an outcome. Here, we focus on statistical abnormality: When multiple variables are necessary for an outcome, people privilege rarer ones as the "actual" cause(s) (Halpern & Hitchcock, 2015; Hart & Honoré, 1959/1985; Hesslow, 1988; Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Kahneman & Miller, 1986).[3] For instance, oxygen is common, but lit matches are rare; hence, people tend to consider the match an actual cause, and relegate the oxygen to a background condition, even if both were necessary to make a fire burn. Similarly, it is rarer for a student to have an engaging college essay than to submit her materials on time; hence, we are likely to consider the essay, rather than the on-time submission, a cause of her acceptance.

The idea that actual causes tend to be abnormal is widespread. It plays an important role in numerous philosophical and legal theories of causation (Collingwood, 2014; Hart & Honoré, 1959/1985; Mackie, 1974), and has begun to be incorporated into prominent formal models of actual causation (Halpern & Hitchcock, 2015). And, like necessity, it is supported by

[2] An important caveat to our analysis is that we focus exclusively on "dependence" notions of actual causation, as opposed to "process" notions (Sloman & Lagnado, 2015; Walsh & Sloman, 2011; Wolff, 2007). On a dependence approach, outcomes somehow *depended* on their causes. The presence of the forest fire depended in some way on the presence of the match, and the theoretical challenge is to understand that dependence relation. In contrast, on a process approach, causes are connected to outcomes via some physical process such as energy transfer (Dowe, 2000). There is substantial debate about to what extent people employ these two approaches (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2015; Lombrozo, 2010; Walsh & Sloman, 2011). We do not engage with this debate. Instead, we show that a particular dependence approach provides an elegant normative explanation for several prominent empirical patterns in actual causal judgments.

[3] People's causal judgments also depend on *prescriptive* abnormality; they privilege variables that violate prescriptive or moral norms (Alicke, 2000; Hitchcock & Knobe, 2009). We return to prescriptive abnormality in the discussion at the end of the paper.

a wealth of psychological evidence (Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard et al., 2017; Kahneman & Miller, 1986; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015).

Our model and our data build very directly on these findings, and so it will be helpful to describe one example from the literature in detail (Icard et al., 2017). Participants were told that a building's climate control system only activates if two employees are in the building at the same time (cf. Kominsky et al., 2015). One morning, two employees, Billy and Suzy, both arrive at 9 AM and the system activates. Participants were then asked both whether Billy and Suzy caused it to activate. Critically though, Icard et al. manipulated whether Billy typically or rarely arrives at 9 AM, and found that people judged him much more of a cause when he *rarely* arrived at that time (Fig. 1). This result exemplifies "abnormal selection" and, along with many similar results, it shows how humans often select rare variables as the actual causes of subsequent outcomes. (We use this causal structure, where two independent variables are each necessary for an outcome, as a simple test case throughout the paper.)

Supersession More recent research identifies a third signature pattern of actual cause judgments: The perceived causal status of a variable also depends on the abnormality of *other* variables in the causal system (Kominsky et al., 2015, cf. McGill & Tenbrunsel, 2000). The pattern has been established most clearly in the two-variable conjunctive structure used above (Figure 1A). People believe that Billy is more of a cause of the climate system activation when he rarely arrives at 9 AM (the abnormal selection effect; Figure 1B). But they also believe that Billy is *less* of a cause when *Suzy* rarely arrives at 9 AM. In other words, in a causal system where two variables are needed to produce an outcome, people believe that one variable is less of a cause when the other variable is rarer. Kominsky et al. (2015) labeled this effect "causal supersession", capturing the idea that causal attribution to one event can be "superceded" by causal attribution to a less normal one. Although only recently characterized, this effect has been replicated in other studies (Icard et al., 2017; Samland & Waldmann, 2016); we also replicate it and expand on it in our behavioral experiments below.

An exception: Overdetermination and preemption In sum, when asked what caused an outcome, people tend to favor variables which were necessary for the outcome, variables which were statistically rare ("abnormal selection"), and variables whose counterparts were *not* statistically rare ("supersession"). Our model will offer a normative justification for these patterns.

There is, however, a well-known exception to these patterns: cases of overdetermination, or its cousin preemption. In these cases, there are multiple events which are each sufficient to produce an outcome. For instance, consider a firing squad with two soldiers, each of whose bullets are guaranteed execute the prisoner alone. Which of their shots caused the prisoner to die? In cases of overdetermination, both events happen simultaneously (e.g. both soldiers shoot at once). In cases of preemption, one event happens first, but the other would have brought about the effect if the first had not (e.g. one soldier shoots before the other, and the prisoner dies before the second bullet hits him).

In these cases, all three effects—necessity, abnormal selection, and supersession—are either absent or reversed. In overdetermination cases, neither event is necessary for the outcome, and yet people judge both causal (Gerstenberg, Goodman, et al., 2015; Gerstenberg, Halpern, & Tenenbaum, 2015; Gerstenberg & Lagnado, 2010; Lagnado, Gerstenberg, & Zultan, 2013; Zultan, Gerstenberg, & Lagnado, 2012). The first soldier is not necessary for the prisoner's death because, if he hadn't shot, the second soldier's bullet would have carried out the execution anyway; and yet the first soldier still seems causal. (Similar logic applies to the preemption case.) Moreover, in cases of overdetermination,

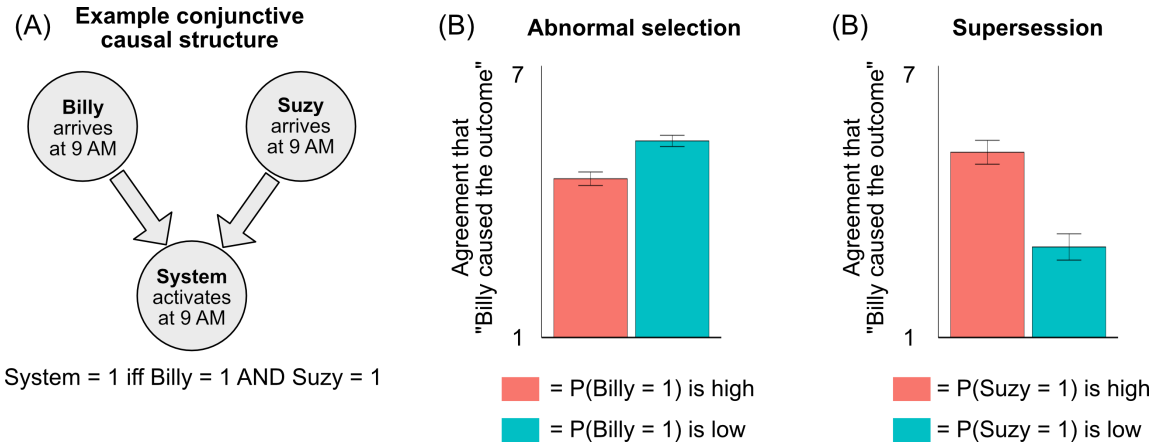


Figure 1: *(A) A conjunctive causal structure, where two variables are each necessary (and jointly sufficient) to produce an outcome. (B) Results from Icard et al. (2017) demonstrating the "abnormal selection" effect. Icard et al. presented participants with this causal structure, and qualitatively manipulated the prior probability of one of the variables. People judge the variable more causal when it is rarer. (C) Results from Kominsky et al. (2015) demonstrating the "supersession" effect. Using the same causal structure, Kominsky et al. manipulated the prior probability of the other variable. People judge the original variable more causal when its counterpart is more common. All results reproduced with permission of the authors. (Note that these authors used slightly different vignettes than the example employed here.)*

abnormal variables are no longer considered more causal, and their causal status no longer depends on the normality of other variables in the system (Icard et al., 2017; Kominsky et al., 2015). (In fact, in cases of overdetermination, abnormal variables are considered less causal; Icard et al., 2017.)

An enormous amount of work has been devoted to understanding these kinds of cases, and any theory of actual causation must ultimately explain them (Hall, 2004; Halpern, 2016; Halpern & Pearl, 2005; Hitchcock, 2009; Lagnado et al., 2013; Lewis, 1973; Sartorio, 2012). We will largely set cases of overdetermination and preemption aside, however, and focus instead on non-overdetermined cases: conjunctive situations (where the relevant variables are all needed to produce an outcome), or disjunctive situations where only one of the potentially sufficient events occurs. Our goal is to provide a normative justification for the three patterns exhibited in non-overdetermined cases—necessity, abnormal selection, and supersession—and then test whether novel judgments in these cases can be predicted by the functional model we propose. Our strategy is to secure some progress in simpler cases, hoping that this will provide new tools to approach more difficult cases later. Following this approach, we will return to the question of overdetermination and preemption in the discussion, considering how our model might be extended to known patterns of judgment in these situations.

More generally, we assume that our model will not provide a perfect fit to all known aspects of causal judgment. This is a natural consequence of its goal, which is not only to mirror the broad shape of human causal judgment, but also to provide a normative justification for its most salient contours. In other words, the model is designed to show that human causal judgment is structured very much like an efficient solution to an important problem. (In Marr's (1982) hierarchy, our model resides at the computational level of analysis.) It is possible that the exact algorithm we derive—no more, and no less—is implemented in the human brain; possible, but very unlikely. More likely, the algorithm we derive is closely related to certain parts of a larger and more complex apparatus for causal judgment. Our

key claim is just this: The functional design of the algorithm matches an important part of the functional design of causal judgment. Specifically, both use counterfactual reasoning over events sampled from experience in order to approximate the general effectiveness of future interventions to bring about some desired outcome.

ii. Existing theories

How do existing theories explain these empirical patterns? Dependence theories of causation (such as ours) traditionally fall into one of three camps: covariation, counterfactual, and interventionist theories.

According to early psychological theories, we judge as causes those variables that co-vary with an outcome; the probability of observing the outcome is higher after observing the cause (Cheng, 1997; Cheng & Novick, 1990; Perales & Shanks, 2007; Spellman, 1997; Ward & Jenkins, 1965). For instance, the probability of a window breaking is higher after a baseball is thrown at it; hence, the baseball was a cause of the window breaking. These "covariation" theories come in many flavors, some of which offer explanations for abnormal selection or supersession (see Icard et al., 2017). (We describe these in more detail with our behavioral experiment, where we compare their empirical predictions to those of our own model.)

One shortcoming of covariation theories is that they do not capture one of the most basic pattern in judgments of actual causation: That causes tend to be necessary for their outcomes (Cheng, 1997). Imagine a rooster that crows every morning before the sunrise. The rooster crowing and the sun rising may covary, but people do not think the rooster caused the sunrise. Presumably this is because the rooster was not necessary for the sunrise—if the rooster hadn't crowed, the sun would still have risen. Covariation-based theories struggle to explain such patterns because they do not involve a representation of necessity.

Due in part to this deficit, two alternative approaches have become popular. These approaches ground causal judgments in the language of counterfactuals (Lewis, 1979). One approach focuses on the similarity between "possible worlds": On this view, a variable caused an outcome if (roughly), in the closest possible world in which the variable didn't take its present value, the outcome didn't occur (Lewis, 1973, 1979; Paul & Hall, 2013).

We build on the second approach, which instead grounds causal judgments in counterfactual *interventions*. On this view, a variable caused an outcome if intervening on the variable would have altered the outcome in some appropriate way (Woodward, 2003). For instance, a student's essay might be considered a cause of their acceptance because, had someone intervened to make the essay unengaging, the student would have been rejected. In much the same way, the rooster is not considered a cause because intervening on the rooster to make it crow would not make the sun rise.

Counterfactual and interventionist approaches are similar, and may in fact be compatible (Pearl, 2000).[4] But the interventionist approach is exciting because it offers a foothold on the "why" question. The insight of the interventionist approach is that causation is connected to *action*—to behaviors like interventions. People are constantly deciding how to act on the world, and causal judgments may be designed to guide choice among potential interventions (Hagmayer & Sloman, 2009; Hitchcock, 2017; Meder et al., 2010; Sloman & Hagmayer, 2006).

iii. The *why* problem: Good interventions and a functional approach to actual causation

What, exactly, is an intervention? An intervention picks a specific variable, severs its ties to any antecedent variables that may have naturally influenced it, and then sets that variable to a specific value.[5] When a scientist

[4] Considering an intervention on a variable, e.g. to fix its value to zero, may be equivalent to considering the closest possible world in which the variable equaled zero. See Halpern (2013) for discussion.

[5] See Woodward (2003) for a more thorough analysis.

injects a rat with caffeine and measures its subsequent energy levels, she has intervened on the rat's caffeination level; when a toddler throws his applesauce on the floor and waits for his mom's reaction, he has intervened on the floor's messiness. According to interventionist theories, these types of actions are central to causal judgment; causes are variables which you can intervene on to alter the probability of effects (Woodward, 2003, 2007).

Interventionist theories offer a natural explanation for the very existence of causal selection—that is, why people foreground some variables as actual causes, while backgrounding others. If, in general, causes are variables on which you can intervene to manipulate effects, then perhaps actual causes are particularly *good* interventions—ones that work well across a variety of situations (Hitchcock, 2012; Lombrozo, 2010; Vasilyeva, Blanchard, & Lombrozo, 2016; Woodward, 2006). An intuitive example comes from Hitchcock and Knobe (2009). Imagine that your paper is sent to a reviewer who believes that there should be no more than three uses of the word "and" per page. Since the paper violated this rule, it gets rejected. When your colleague asks why your paper got rejected, are you more likely to say it was because of your promiscuous use of conjunctions, or because the editor sent it to a crazy reviewer? Probably the latter. Hitchcock and Knobe argue that this is because, even though intervening on either variable would have rectified the problem in *this* situation, in general it is a better strategy to blacklist this reviewer (or avoid this editor, or stop submitting to this journal) than to minimize conjunctions in your journal articles forever after. People highlight certain variables as actual causes because they recognize that intervening on those variables is, in general, a better way to manipulate the relevant outcome.[6]

Yet, there is a subtle ambiguity in this position. How, exactly, should the relationship between "good interventions" and "actual causes" be understood? We contrast two possibilities: Does existing knowledge of good interventions guide causal selection, or does causal selection build future knowledge of good interventions?

Beginning with the first possibility, perhaps people *first* identify effective interventions, and *then* label those variables actual causes. In the example above, you might first judge that blacklisting the reviewer is an effective strategy for future paper submissions, and then label the reviewer as the actual cause of your current rejection. This runs into a serious difficulty, however. This interpretation undermines the original intent of our argument: to provide a functional account of actual causation (i.e. to answer the "why" question). If some representation of "good interventions" exists and is used to select among the causes of past events, then this very same representation of "good interventions" should be *directly* queried when planning for the future. In other words if you already know the good interventions that get papers accepted, then judging the reviewer the cause of your current rejection serves little purpose for your future plans. When you are planning how to get your next paper accepted, don't worry about the past—just choose one of the good interventions! But if you have already solved the problem of finding optimal future interventions, why exactly are you bothering to identify optimal past ones?

As Hitchcock (2017) puts it, there is an apparent mismatch between the orientation of causal judgment (towards the past) and planning (towards the future). Actual causation is "backward looking" (Hitchcock, 2017); it is anchored in the idiosyncratic details of the past event whose cause is being identified. In contrast, knowing which interventions tend to bring about an event in general is a forward-looking problem—the problem of decision-making. What purpose, then, is served by judging *specific, past* causes against the yardstick of *general, future* utility (Lombrozo, 2006)?

In light of this challenge, we focus on the second possible relationship between these variables (Figure 2). We propose that actual causes

[6]This idea is supported by the fact that, when identifying necessary causes, people sometimes focus on variables which were "controllable" (Girotto, Legrenzi, & Rizzo, 1991); but see Mandel and Lehman (1996) for an alternative perspective.

are not selected on the basis of some prior notion of "good intervention". Rather, actual causal judgments are the very inputs into the process of discovering good interventions. The causal judgments rendered in many specific past episodes gradually accumulate evidence about the general effectiveness of future interventions. We call this approach the "Sample-based Approximation Method for Predicting the Likelihood of Effectiveness" — or the SAMPLE approach, for short. The SAMPLE approach illuminates the role of actual causation in approximate planning, and illustrates how decision-makers can identify effective interventions, even in complex scenarios where precise computation is intractable. (We mean SAMPLE to describe our general proposal regarding the functional design of actual causation judgments, not the specific mathematical form of the algorithm we define below.)

Our next challenge is to specify an algorithm that delivers actual causal judgments of this form. In the remainder of the paper, we formalize this approach in the language of structural equation modeling and, focusing on a simple but common decision problem, we show what form actual causal judgments would have to take in order to accumulate evidence about effective interventions. This form turns out to capture the empirical patterns identified above: In order to tractably approximate intervention effectiveness, actual causal judgments should be rooted in necessity, and should exhibit both abnormal selection and supersession effects. We thus offer a precise normative justification for those patterns. Finally, we test predictions of the SAMPLE approach in a novel behavioral experiment.

III. Formal model of effective intervention and actual causation

We propose that actual causal judgments accumulate information about the effectiveness of interventions. To make this proposal precise, we will develop it in a specific decision problem. Imagine that a person wants to bring about an outcome E, such as starting a fire, getting a paper accepted, or getting a child into college. The person has a model of the causal system in which E is situated.[7] Moreover, the person can intervene on the system by turning on a single variable, C_i, from a set of candidate variables $C_1, ..., C_K$. (By "turning on" a variable, we mean setting its value to 1.) For instance, the admissions officer trying to get a child into college could hire a writing coach, pay for piano lessons, move to a better school district, and so on. Which one should she choose?

To succeed at this decision, the person must assess the "effectiveness" of the potential interventions $C_1, ..., C_K$. We will leverage this decision problem to produce a precise definition of "intervention effectiveness", and present an algorithm in which actual causal judgments cumulatively approximate the effectiveness of interventions. To implement this approximation, actual causal judgments must take a certain form, which we will unpack, and then explain why this reproduces the empirical patterns of necessity, abnormal selection, and supersession.

Focusing on this well-defined decision problem has a number of advantages. It isolates the role of intervention effectiveness; the person's choice depends only on the probability that potential interventions will produce a desired outcome. This simplification allows us to highlight the proposed function of actual causal judgment—estimating the effectiveness of potential interventions—while sidelining extrane-

[7] Throughout our proposal, we will assume that the person making judgments and decisions already has an accurate causal model of her environment. This assumption allows us to sideline the difficult problem of causal induction (how people infer causal structures from sparse data; Bramley, Dayan, Griffiths, & Lagnado, 2017; Bramley, Gerstenberg, Mayrhofer, & Lagnado, accepted; Coenen, Rehder, & Gureckis, 2015; Griffiths & Tenenbaum, 2005, 2009; Meder et al., 2010; Rottman & Hastie, 2013; Stephan & Waldmann, 2016; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003), and focus exclusively on causal selection — why people privilege some causes as "actual", even when they know the structure of their environment (as well as what actually happened). This model need not be explicit, or accessible to conscious deliberation; the only requirement is that people can perform the computations described below (e.g. assessing necessity, computing base rates).

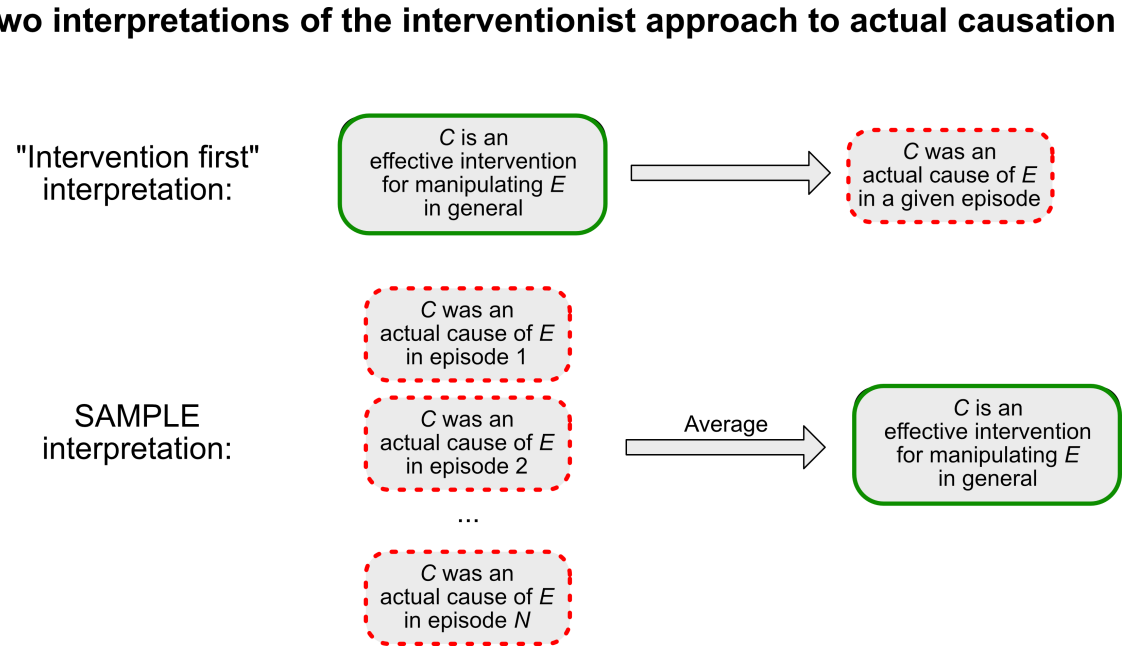


Figure 2: *Schematic depiction of two possible relationships between actual causes and effective interventions. C represents the potential cause (e.g. the match); E represents its effect (e.g. fire).*

ous decision processes (like selecting a goal, or incorporating action costs). The simplicity of this decision problem also affords a rigorous definition of intervention effectiveness, and allows us to derive a tractable approximation algorithm with precise empirical implications.

Our focal decision problem makes a number of simplifying assumptions. For instance, it assumes that all variables are binary — either 0 ("off" or absent) or 1 ("on" or present). This assumption, shared with other formal treatments of causation (Halpern & Pearl, 2005; Hitchcock, 2001; Icard et al., 2017), makes our analysis tractable, but may often be unrealistic. We also assume that a person is considering turning *on* a variable in order to *produce* an outcome. This assumption allows us to avoid tricky questions about variable absences and actions/omission distinctions. In the discussion we return to both of these simplifying assumptions and mention a few others, explaining how our model could be expanded to encompass more complex cases.

A virtue of our simple decision problem is that, within its bounds, we find a remarkably accurate fit between our algorithm and people's actual causal judgments, as demonstrated in our experiment below. One possibility is that actual causation is designed to solve just the exact problem we consider here—identifying optimal interventions—and no other. More likely, however, we have isolated one important function of causal judgment among several. In other words, we propose that the underlying principles identified here apply in more complex decision scenarios, and that this simple test case accurately characterizes important aspects of just one of the problems that actual causal judgments are designed to solve. An important goal for future research is to explicitly apply the SAMPLE approach to more general decision problems.

i. Defining intervention effectiveness

In the simple decision problem described above, the decision-maker's job is clear: Roughly, she should choose a variable such that, after turning that variable on, there is a high probability that the outcome is on—given

everything that she knows about her current situation (Hagmayer & Sloman, 2009; Sloman & Hagmayer, 2006). For instance, if an arsonist knows that the forest ground is currently dry, this knowledge could alter the predicted downstream effects of lighting a match; and so she should consider it when computing the probability of a subsequent forest fire.

To perform these computations, the decision-maker must have a model of the causal system she is facing—not just of the potential intervention (e.g. the match), but also of the other variables (the ground dryness, the oxygen in the air) that influence the desired outcome. Structural equation modeling (Pearl, 2000) offers a formal language for describing causal systems, and thus allows us to precisely define the effectiveness of potential interventions.

Structural equation modeling approach A structural equation model is a formal description of a causal system that allows a decision-maker to compute the downstream effects of interventions (Pearl, 2000). It is composed of three parts: a set of exogenous variables, a set of endogenous variables, and functions that relate them to each other. The exogenous variables represent all the "unknowns" of the causal system—the elements whose antecedents the decision-maker does not want to explicitly model. For instance, one might create a structural equation model of forest fires (Figure 3A) where whether or not an arsonist is present is an exogenous variable: the decision-maker does not model the variables that determine it. Instead, for convenience, she simply assigns a prior probability to it (e.g. there might be a 10% chance that, in any observation of the system, an arsonist is present).

In contrast, the endogenous variables represent the elements that the decision-maker *does* want to model. The values of the endogenous variables are determined by the other variables in the causal system. For instance, in our toy example, there is a forest fire if and only if there is oxygen, dry ground, and a lit match. This relationship can be captured by describing the variable "forest fire" (FF) as a function of the variables "oxygen" (O), "dryness" (D), and "lit match" (LM). In this case, $\texttt{FF} = min(\texttt{O}, \texttt{D}, \texttt{LM})$; O, D, and LM must all equal 1 for FF to equal 1. Of course, the decision-maker does not typically know the values of all the exogenous variables. Hence, to compute the probability that an *endogenous* variable takes on a value (e.g. $Prob(\texttt{FF} = 1)$), she simply identifies the settings of the *exogenous* variables that would produce such a value (e.g. $\texttt{O} = 1, \texttt{D} = 1, \texttt{A} = 1$), and computes the probability of those settings. This technique allows her to simultaneously represent the deterministic forces that produce an event, and also the uncertainty surrounding its actual occurrence (i.e., a probability distribution).[8]

Structural equation models support two operations that are essential for defining the effectiveness of an intervention: intervening and conditioning. Intuitively, an intervention selectively alters the value of a single variable, severing its ties to antecedent variables that may have influenced it. Structural equation models formalize this with the notion of a *mutilated model*. Intervening to set a variable V to value v in structural equation model M, the decision-maker considers a derivative model $M_{V=v}$ in which the function that determines V is replaced with $V = v$. (This process is often referred to as applying the "do" operator; see Pearl, 2000; Sloman & Lagnado, 2005.) Then, to determine the effects of this hypothetical intervention, the intervener computes the prob-

[8] You might object that this setup is implausible. Endogenous variables, the objection goes, are never completely determined by the other variables in the causal system. Rather, their values are probabilistic, and the settings of the other variables simply change their conditional probabilities. (This is the approach taken, for example, by causal Bayes nets.)

The response is that this apparent stochasticity is only a reflection of our limited model. The reason forest fires do not always happen in the presence of oxygen, dry ground, and a lit match is that there are other variables that must also take on certain values. In theory, if we include enough variables (or a "catch-all" variable representing the ones we don't explicitly enumerate), then the endogenous variables become completely determined by the rest of the causal system. This is called the "Laplacian assumption" (Pearl, 2000). (Another possibility is that the causal models are deterministic, but there is uncertainty about which model is correct; see Glynn, 2017; Halpern, 2016).

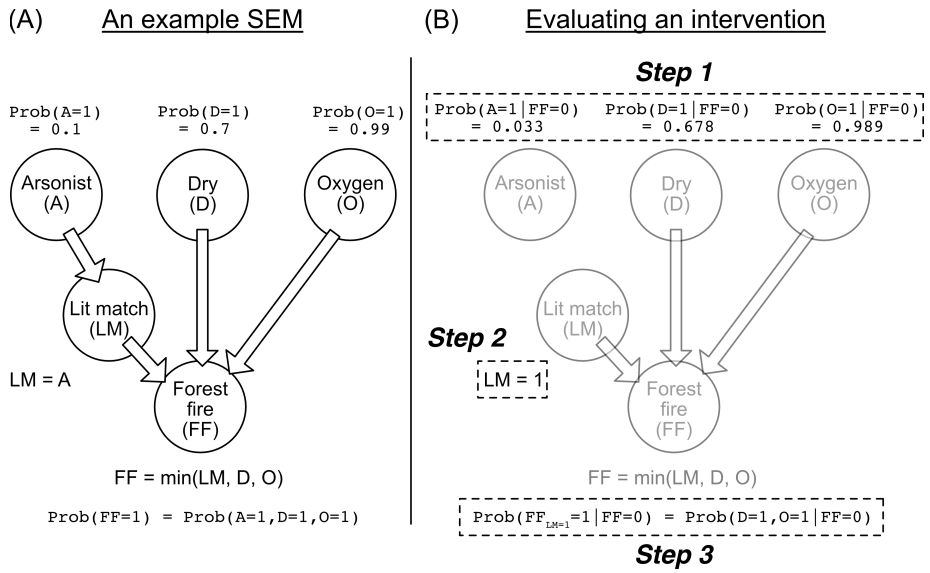


Figure 3: *(A) An example structural equation model of forest fires. Conceptually,* $A = 1$ *indicates that there is an arsonist;* $D = 1$ *indicates that the ground is dry;* $LM = 1$ *indicates that someone lit a match;* $O = 1$ *indicates that there is oxygen; and* $FF = 1$ *indicates that there is a forest fire.* A*,* D*, and* O *are exogenous;* LM *and* FF *are endogenous. In this toy model, we assume that there is a lit match if and only if there is an arsonist, and there is a forest fire if and only if someone lights a match, the ground is dry, and there's oxygen in the air. (Conventionally, exogenous variables are not depicted in graphical representations – but we show them here for exposition.). (B) Procedure for computing the probability of fire after lighting a match, conditioned on* FF *currently being zero. First, you update the probabilities of the exogenous variables to account for* $FF = 0$*. Second, you intervene to fix* $LM = 1$*. Third, you sum the conditional probabilities of the exogenous variable settings that would, post-intervention, produce* $FF = 1$*.*

abilities that other variables would take certain values in this mutilated model. For instance, to compute the probability that there would be a fire if you lit a match, we compute the probability that $\texttt{FF} = 1$ in the mutilated model with $\texttt{LM}$ fixed to 1 (notated $Prob(\texttt{FF}_{\texttt{LM}=1} = 1)$).

The second operation, conditioning, involves updating the probabilities in the causal model based on knowledge of variable values. For instance, one could compute the probability that there would be a forest fire if you lit a match—conditioned on your knowledge that, currently, there is no fire. Formally, we notate this $Prob(\texttt{FF}_{\texttt{LM}=1} \mid \texttt{FF} = 0)$. This operation proceeds in three steps (Pearl, 2000; Figure 3B). First, the decision-maker computes the conditional probabilities of the exogenous variables given her knowledge of the system's current state (i.e. $Prob(\texttt{A} = 1 \mid \texttt{FF} = 0)$, $Prob(\texttt{D} = 1 \mid FF = 0)$, etc.). Second, she mutilates the model to simulate the intervention (fixing $\texttt{LM} = 1$). Finally, she computes the probability of the outcome ($\texttt{FF} = 1$) in the mutilated model, using the conditional probabilities from step one. In this way, structural equation models allow her to compute the effects of interventions while incorporating her knowledge of specific situations.[9]

Situation-specific effectiveness With this apparatus in place, let's return to our paradigmatic scenario. A decision-maker is interacting with a causal system with variables, V. One variable, $E \in V$, is a desired outcome. Our decision maker must pick one of k variables, $C_1, ..., C_k \in V$, on which to intervene. Moreover, she knows something about the situation; she knows the value of some subset of the vari-

[9] Pearl (2000) calls this process "counterfactual inference". For a more formal treatment of it, and of structural equation modeling in general, see Pearl (2000).

ables $V_{known} \subseteq V$.

In this setup, the optimal intervention is clear: The decision-maker should choose C_i that maximizes the post-intervention probability of $E = 1$ conditioned on $V_{known} = v_{known}$. Formally, she should choose C_i that maximizes $Prob(E_{C_i=1} = 1 \mid V_{known} = v_{known})$. Call this the *situation-specific effectiveness* of intervention C_i. Consider a more concrete example: if a parent is trying to get her child into college and knows that the child is good at math but not at writing, then the situation-specific effectiveness of hiring a writing coach would be $Prob(\texttt{GetAccepted}_{\texttt{HireWritingCoach=1}} = 1 \mid \texttt{GoodAtMath} = 1, \texttt{GoodAtWriting} = 0)$.

For a person with unlimited computational resources facing our simple decision problem, this situation-specific approach would be an ideal notion of an effective intervention. The agent's only goal is to bring about the outcome $E = 1$, and this measure identifies the interventions most likely to achieve that goal (given the agent's knowledge of the situation). Unfortunately, in any causal system of real-world complexity, computing this probability involves non-trivial costs, and will often be intractable at decision time. Specifically, it requires evaluating all possible joint settings of the unknown exogenous variables—in plain English, computing the probability of all conceivable states of affairs consistent with one's limited knowledge of the world. For instance, for the parent to compute the situation-specific effectiveness of hiring a writing coach, she would have to enumerate all settings of the unknown exogenous variables under which, after hiring the coach, her child would get accepted, and then sum their conditional probabilities. She'd have to consider the probability that the coach is bad, that the child follows the coach's instructions, that the child doesn't die in an accident, and so on. In a system with 10 binary exogenous variables whose values are unknown, she would have to, in the worst-case scenario, evaluate over a thousand possible scenarios.[10]

In spite of these difficulties, people are often required to—and *do*—identify effective interventions quickly (Bramley et al., 2017; Hagmayer & Sloman, 2009; Markant & Gureckis, 2014; Sloman & Hagmayer, 2006). How could this be accomplished?

General effectiveness as a decision prior This problem could be addressed if people represented the "general" effectiveness of interventions, averaging across situation-specific details. Many have proposed a similar idea—that a useful representation of good interventions should be "robust" or "portable", rather than being tethered to a highly specific state of affairs (Franklin & Frank, 2018; Hitchcock, 2012; Lombrozo, 2010; Vasilyeva et al., 2016). For instance, a person might know that lighting a match is, in most situations, an effective way to start a fire—even if there are some rare situations where it is ineffective (e.g., in strong winds). This representation would serve as a kind of "prior" on the true effectiveness of an intervention in a particular situation, which would be useful for two reasons. If someone did not have time to compute the effectiveness of each potential intervention conditioned on their situation-specific knowledge, they could quickly choose an intervention which was generally effective. And if someone did have time, the "general effectiveness prior" would point to good places to start the more difficult computation. (Evaluating the situation-specific effectiveness of lighting a match is likely a better investment of computational resources than evaluating the effectiveness of pumping oxygen into the air.) Indeed, recent AI approaches to decision-making in large choice spaces often crucially depend on having good priors that guide choice evaluation, as in Monte Carlo tree search (e.g. Silver et al., 2016).

We don't have to completely abandon *all* situation-specific knowledge, however. Across most settings where the decision-maker has a goal of producing an outcome, there is at least

[10] There is another practical obstacle to identifying optimal interventions: How do people know which variables to include in their causal model? The optimal intervention will often depend on which variables are included. We will not address this issue here, but see Halpern and Hitchcock (2011).

one constant: The outcome does not currently obtain. When someone is trying to start a fire, they know there is not currently a fire; when someone is trying to win an election, they know they have not already won. (If the outcome already obtained, then the person would not be employing decision-making machinery on how to obtain it!)[11] In other words, if we're trying to identify generally effective interventions in contexts where a person is trying to bring about $E = 1$, we can safely condition on $E = 0$. (For an in-depth justification of this approach, see section *i* in the Appendix.)

This leads us to our definition of an effective intervention. The general effectiveness of intervening on C for bringing about E is $Prob(E_{C=1} = 1 \mid E = 0)$.[12]

Definition 1. The **general effectiveness** of intervening on C to produce E, denoted $GE(C \rightarrow E)$, is $Prob(E_{C=1} = 1 \mid E = 0)$.

To reiterate, we propose that judgments about whether C was the actual cause of E help approximate the general effectiveness $GE(C \rightarrow E)$, and that $GE(C \rightarrow E)$ can help a person make efficient decisions in future settings when she desires to bring about E. For instance, if the parent can use causal judgments to infer that hiring the writing coach is generally an effective intervention, then, in a time crunch, she can quickly choose to do it; or, if she has time, she can focus on computing the situation-specific effectiveness of hiring the coach, and avoid wasting resources on less generally useful strategies.

An important qualification to this analysis is that the effectiveness of an intervention is fundamentally a question of decision making and planning, not causality. Yet, to model it, we employ the language of structural equation modeling—a framework designed to capture causal reasoning. This is a non-standard usage of the SEM framework. Future work may improve on our model by casting it in a more decision-oriented framework like Markov decision processes (Sutton & Barto, 1998), which would support the analysis of more complex decision problems (e.g. involving action costs, or sequential decisions). However, analyzing the effectiveness of interventions in the SEM framework is an important starting point. On our proposal, causal reasoning and decision making are intimately linked (Hagmayer & Sloman, 2009; Sloman & Hagmayer, 2006). Structural equation models (and their cousin, causal Bayes nets) are the dominant framework for causal reasoning in cognitive science (Pearl, 2000; Sloman & Lagnado, 2015), and, as our analysis shows, are powerful enough to capture meaningful claims about the decision process (Sloman & Hagmayer, 2006). Casting our model in the language of SEMs allows us to make solid contact with both the causal reasoning and decision making literatures.

ii. Actual causal judgments approximate general effectiveness

The general effectiveness of interventions can be approximated from experience. After all, each specific situation a person has experienced is a *sample* from the "general" distribution over situations. Put somewhat differently, each time a person observes a causal system, they are observing a snapshot of the variable settings. If these snapshots are randomly sampled, they can replace the computationally taxing process of enumerating states of affairs. For instance, instead of evaluating all possible college application situations, the admissions officer can rely on past applicants as samples.

Crucially, however, a person must *do* something with the samples they observe; they must perform some computation on each observation. We propose that judgments of actual cau-

[11] A potential counterargument to this approach is that, sometimes, people try to bring about an outcome, even though they are unsure of whether the outcome currently obtains. For instance, a general on the Western front might take steps to win a war even if he is unsure whether, because of the unknown result of a decisive battle on the Eastern front, the war has not already been won. But these situations are relatively uncommon. Moreover, the general might condition on the war not yet being won anyway—because that's the situation in which his actions would matter most.

[12] In previous research, this value has been referred to as the "probability of enablement" (Pearl, 1999). We do not use this name to avoid confusion with the verb "enable".

sation are the output of this computation, and when averaged across many samples, they approximate $GE(C \rightarrow E)$. Our aim is specify an efficient algorithm that renders actual causal judgments which implement this approximation.

Suppose a decision-maker has a series of observations of the system $u_1, ..., u_k$, where each u_i is a joint setting over the system's variables (e.g. in u_7, there was an arsonist, the ground was dry, and there was oxygen in the air). Suppose that she restricts her analysis to members of this set where $E = 1$—i.e. when the outcome of interest is present. (Assume these experiences were randomly drawn from the system's prior distribution, conditioned on $E = 1$.) Then, to approximate the general effectiveness of C for E, she should judge (1) whether $C = 1$ was necessary to produce the outcome in each case, (2) if it was, apply a correction term that weights rare causes more heavily, and (3) average those judgments together.

We propose that a key function of actual causal judgments is to implement the first two steps of this procedure. To judge the extent to which C was an actual cause of E in a particular sample u_i, you determine whether C was necessary for E in u_i, and, if it was, weight it according to (among other things) the rarity of C. These judgments then assist in decision-making, because the person can simply maintain a running average of them to estimate $GE(C \rightarrow E)$.

Formally, we will use the notation $AC_{SAMPLE}(C \rightarrow E, u_i)$ to denote the extent to which C was an actual cause of E in observation u_i. Definition 2 presents our analysis of actual causation.

Definition 2. On our model, the extent to which C is an **actual cause** of E in a situation with exogenous settings u_i is:

$$AC_{SAMPLE}(C \rightarrow E, u_i) = I(E_{C=0} = 0 \mid U = u_i) \frac{Prob(C = 0)}{Prob(C = 1)} \frac{Prob(E = 1)}{Prob(E = 0)}$$

where $I(E_{C=0} = 0 \mid U = u_i)$ is an indicator function equaling 1 if turning C off would have turned E off in u_i, and 0 otherwise. In other words:

$$AC_{SAMPLE}(C \rightarrow E, u_i) = \begin{cases} \frac{Prob(C=0)}{Prob(C=1)} \frac{Prob(E=1)}{Prob(E=0)} & \text{if } C \text{ was necessary for } E \\ 0 & \text{otherwise} \end{cases}$$

If actual causal judgments are guided by this formula, then the general effectiveness of C for E can be estimated by simply averaging the actual causation judgments from each situation. A proof of this proposition is provided in section *ii* of the Appendix.

Proposition 1.

$$GE(C \rightarrow E) \approx \frac{1}{k} \sum_{i=1}^{k} AC_{SAMPLE}(C \rightarrow E, u_i)$$

As we discuss below, there are certain conditions, e.g., the "no-confounding" condition, which must hold for this approximation to work. Before turning to these details, however, we first want to give an intuition for our proposal in a more concrete example. Consider again the admissions officer. If her actual causal judgments of past applicants align with Definition 2, our suggestion is that she will judge engaging writing an actual cause of past applicants' acceptances to the extent that:

1. A given past applicant's acceptance depended on her engaging writing (necessity), i.e. C necessary for E in u_i,
2. Engaging writing is rare, i.e. $\frac{Prob(C=0)}{Prob(C=1)} \gg 0$, and
3. Acceptance is common, i.e. $\frac{Prob(E=1)}{Prob(E=0)} \gg 0$ (which, as described below, predicts super-session).

Having made these actual causal judgments in the cases of past applications, she can then approximate the general effectiveness of hiring a writing coach by simply averaging her past judgments of actual causation. Such an algorithm is valuable because it allows her to estimate the general effectiveness of interventions in a tractable manner. She can then use

that knowledge to guide her search for effective interventions in planning (e.g., for her own children), either by choosing an intervention with high general value, or by devoting situation-specific planning especially towards interventions with high general value.

Why does this approximation work? We answer this question in detail in the next section. But as an overview: We want to know the general effectiveness of getting from $E = 0$ to $E = 1$ by turning on C. Clearly, this could be estimated by just computing the average effectiveness of C in past, randomly sampled episodes. One would simply ask: For what proportion of actual past episodes in which $E = 0$ did I successfully get $E = 1$ by turning on C? (This is not a form of causal judgment, of course, but something more like reinforcement learning; Sutton & Barto, 1998). This simple approach falters, however, when such episodes, or opportunities to intervene in them, are rare.

We propose that actual causation implements an alternate approximation—one that does not depend on trying actual interventions when an effect is absent, but that instead enables learning from counterfactual simulations over episodes where the effect is present. The existence of an episode where E was 0 and turning on C worked—call it episode u_i—implies a corresponding possible episode u_i' in which $E = 1$ and the agent *could* have turned C *off* in order to get $E = 0$. Actual causation judgments use the u_i' episodes, where turning C off would have turned E off, to learn about the u_i episodes, where turning C on would turn E on. It is as if we are mentally flipping a light switch on and off, watching a specific room go light and dark. If flipping C on brings a specific room u_i from darkness to light, then flipping C off must bring a corresponding room u_i' from light to darkness. Establishing the correspondence of such rooms—or states u_i and u_i'—is the job of actual causal judgments.

But we are not just trying to establish the existence of such rooms—we want to know their probabilities of occurrence, which tells us how often C is a good point of intervention. Of course, it will not do to simply compute the probability that episodes like u_i' naturally arise, and then treat this as an estimate of the probability of episodes like u_i naturally arise. Although the *existence* of any u_i' implies the *existence* of a corresponding u_i, their relative *probabilities of occurrence* will usually differ. Thus, in order to compute the probability of episodes like u_i from samples of episodes like u_i, we need to perform a statistical correction for their relative probabilities of occurrence. This is accomplished with the correction term $\frac{Prob(C=0)}{Prob(C=1)}\frac{Prob(E=1)}{Prob(E=0)}$. Having done this, we can accumulate evidence from episodes in which the effect was present, and could be eliminated by turning C off, in order to estimate the probability of episodes in which the effect was absent, and could be produced by turning C on.

IV. Unpacking the definition of actual cause

In this section, we further develop the intuition for why actual causal judgments of this form can approximate the general effectiveness of interventions, and how they provide a normative explanation for the empirical patterns of actual causal judgments described above. (For a formal proof that actual causal judgments can approximate $GE(C \rightarrow E)$, see section *ii* in the Appendix.)

The core of our proposal is in Definition 2—the form that actual causal judgments take to approximate intervention effectiveness. There are two parts to the definition: the necessity criterion (whether turning off the potential cause would have turned off the outcome), and the correction term. We take up each in turn.

i. Actual causes should be necessary

According to Definition 2, if actual causation were designed to approximate general intervention effectiveness, then the first step in determining actual causes would be to judge whether "turning off" the potential cause—i.e. intervening to set $C = 0$—would have turned off the outcome. For instance, to determine whether the lit match caused the forest fire,

a person would judge whether intervening to prevent the match from being lit would have prevented the forest fire.

This corresponds to the standard necessity criterion: Judging whether turning off C would have turned off E is a way of assessing whether C was necessary for E. What is promising about this connection is that starting only from normative assumptions about general effectiveness, we can capture the ubiquitous finding that necessity is central to judgments of actual causation.

At the same time, this result might seem puzzling. In our framework, the general effectiveness of an intervention is about whether turning a variable *on* (e.g. lighting a match) would produce an outcome when the outcome is absent—but it's being estimated by considering whether turning a variable *off* (e.g. snuffing out a match) would have prevented an outcome when the outcome is present. Put differently, the definition of $GE(C \rightarrow E)$ involves setting $C = 1$ when $E = 0$, but the definition of $AC_{SAMPLE}(C \rightarrow E, u_i)$ requires people to imagine setting $C = 0$ when $E = 1$. Why?

The first thing to note is that our approximation algorithm is not unique; the general effectiveness of an intervention could be estimated other ways. People *could* estimate $GE(C \rightarrow E)$ more directly by determining whether turning on the variable (e.g. lighting a match) would produce the outcome in situations where the outcome is absent. In other words, instead of relying on judgments of necessity, people could rely on judgments of sufficiency. In fact, at times, people probably do employ this alternative approximation method. Any time a person considers whether turning on a variable would produce an outcome in a given situation, they can use that computation to inform their future decision making more generally. Sufficiency judgments likely play a role in assessing the effectiveness of interventions, but we would typically describe this as a form of planning, rather than causal judgment.[13]

While it is intuitive that sufficiency judgments may play a role in decision making, the key insight of our proposal is that necessity judgments—of the type employed in actual causation—can also play a role, and indeed potentially a very large one. As long as the causal system satisfies certain conditions described below, and after applying the correction term specified in our algorithm, judging whether turning off a variable would have turned off an outcome provides an unbiased estimate of that variable's ability to produce the outcome.

Why do necessity judgments work? Why does this conversion from necessity to sufficiency—from the probability of "worsen writing, prevent acceptance" to the probability of "improve writing, get accepted"—work? It works because any set of background factors (i.e. exogenous variable settings) that makes a variable necessary (when it is on) is identical to the set that makes a variable sufficient (when it is off). Imagine that a student would not have been accepted had she not written an engaging essay (the essay was necessary for her acceptance). This fact implies that, in an identical situation *minus her engaging essay* (and therefore minus acceptance), writing a good essay would have been sufficient to bring about a victory. Necessity and sufficiency identify the same situations, with only the focal variable's setting flipped.

It follows, then, that necessity judgments will provide information about the ability of a variable to produce an outcome (its intervention effectiveness)—so long as the probability of the relevant background factors does not depend on the setting of the focal variable. (A variable is "relevant" here if it affects the post-intervention probability of the outcome.) Judg-

[13] This alternative approximation algorithm would be simple. Suppose we have observations of the system $z_1, ...z_k$, which are randomly sampled from the prior over exogenous variables, conditioned on $E = 0$. In other words, we observed a series of instances where there was no fire, or when no paper was accepted. To estimate $GE(C \rightarrow E)$, we could simply determine whether $C = 1$ (e.g. lighting a match) would have been sufficient to produce $E = 1$ in each case z_i, and average those judgments together. That is: $GE(C \rightarrow E) \approx \frac{1}{k} \sum_{i=1}^{k} I(E_{C=1} = 1 \mid Z = z_i)$. This alternative approximation could then be averaged with the results of the necessity approximation to obtain a better estimate of $GE(C \rightarrow E)$.

ing when the essay was necessary provides an unbiased estimate of when a good essay would be sufficient, so long as the probability of the relevant background factors (e.g. the college's admissions rate, the student's extracurriculars, etc.) does not depend on the status of the essay.

A commonly employed "no-confounding condition" guarantees that the situations which make a variable necessary when it is on will be equally probable when the variable is off. The technical definition, modified for our context, is as follows.

Definition 3. The **no-confounding condition** holds for a potential cause/effect pair (C, E) if and only if:

$$Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = 0) = Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = 1)$$

In other words, the effect on the outcome E from intervening on the potential cause C must be independent of whether C is on or off. (Equivalently, the variable settings that set the backdrop for intervening on C are equally probable when C is on or off.) This technical criterion can be intuitively represented by a graphical criterion. The no-confounding assumption is guaranteed to hold if C and E have no common ancestor—if there is no upstream variable that influences both the potential cause and its effect.

In light of the previous discussion, it is easy to see why the no-confounding assumption supports a conversion from necessity to sufficiency. It guarantees that the background factors which make a variable necessary when it is on are equally likely to occur when the variable is off. The only difference between the necessity-situations and sufficiency-situations are the settings of the focal variable C and its outcome E—and these get accounted for in the correction term $\frac{Prob(C=0)}{Prob(C=1)}\frac{Prob(E=1)}{Prob(E=0)}$, which we discuss below. (For a technical description of the no-confounding assumption's role in our model, see section *ii* in Appendix.)

How plausible is this assumption in real-world causal systems? Though rarely perfectly fulfilled, the no-confounding assumption is plausible enough to be widely adopted in practical fields like epidemiology and law (Pearl, 2000), and is a weaker assumption than those made in popular psychological models (Cheng, 1997; Luhmann & Ahn, 2005). We examine the no-confounding assumption further in the general discussion.

Why use necessity judgments? In sum, if a causal system satisfies the no-confounding assumption, then necessity judgments can be used to approximate intervention effectiveness. We have already noted that this need not be the only way of approximating intervention effectiveness—a sufficiency approximation (i.e., imagining $C = 1$ when $E = 0$) is an obvious alternative. But it is natural to suppose that we draw upon multiple approximation methods: Real-world causal systems are complex, and the more information a person can harness to strengthen their approximation, the better.

There are, however, two reasons that the necessity approximation might be a particularly rich and efficient method. First, when people have a goal of obtaining some outcome $E = 1$, it is typically the case that the presence of the outcome is rare, compared to its absence. People observe many, many more instances with no fire than instances with fire. As a consequence, it would be quite difficult to judge sufficiency in all the instances where $E = 0$. (Imagine asking yourself, every moment of every day, which variables would be currently sufficient to produce a fire.) If similar estimates can be obtained either way, then it is computationally economical to focus on the relatively few instances where $E = 1$.

Second, the structure of many important causal systems might make it easier to assess necessity than sufficiency. Often, when an outcome occurs, people find it easy to generate many variables that were necessary for it; but when an outcome doesn't occur, it is more difficult to generate variables that would have been sufficient to produce it. For instance, imagine that a basketball player shoots a 3-pointer. It is easy for even a casual fan to think of ways he could have missed—if his arc had been a little

lower, if the defending player had jumped a little higher, if his knee had given out, and so on. But, if a player is missing his 3-pointers, it may be difficult to describe exactly what he should change to sink them. If it were all that easy, basketball coaches would be out of a job.

Of course, there are also many cases where sufficiency is intuitively easier to assess than necessity. An important task for future research is to identify the characteristics of causal systems that make necessity or sufficiency more salient. For now, we note that necessity judgments seem salient in enough cases that decision makers would likely profit from a cognitive tool—actual causation—dedicated to harvesting the information they provide.

In sum, by judging whether a variable was necessary for an outcome when the outcome was present, people can determine whether a variable will produce an outcome when it is absent. This result provides a normative justification for the widespread empirical finding that necessity is important for actual causation. Critically, however, this approximation only works if people apply an appropriate correction term. We now turn to discussing this correction term, show why it is important, and show how it provides a surprisingly tight fit with the two other empirical patterns we noted at the outset: "abnormal selection" and "supersession".

ii. The correction term

The second part of our definition of actual cause (Definition 2) is the correction term: If C was necessary for E, then the extent to which it was an actual cause is:

$$\frac{Prob(C = 0)}{Prob(C = 1)} \frac{Prob(E = 1)}{Prob(E = 0)}$$

This term allows necessity judgments when $E = 1$ to accurately approximate the general effectiveness of potential interventions when $E = 0$.

There are three important features of this correction term. The first is that it makes actual causal judgments continuous, not discrete; in other words, one variable can be more of an actual cause than another. This accords with a wealth of empirical evidence that people, in fact, treat actual causation as graded (Danks, 2017; Gerstenberg, Goodman, et al., 2015; Gerstenberg et al., 2017; Halpern & Hitchcock, 2015; Kominsky et al., 2015).

Actual causes should be abnormal The second feature is that it weighs rare variables more heavily (via $\frac{Prob(C=0)}{Prob(C=1)}$). As discussed above, this also accords with a wealth of empirical evidence that people favor rare causes over common ones—the "abnormal selection" effect (Hilton & Slugoski, 1986; Hitchcock & Knobe, 2009; Icard et al., 2017). There are two reasons that abnormal causes are weighted more heavily. The first, captured by the $Prob(C = 0)$ term in the numerator of Definition 2, follows from a basic fact: Turning a variable "on" cannot be a good intervention in any situation in which it is already "on". Thus, in order for turning a variable on to be a good intervention in general, the variable must generally be "off".[14]

Consider again a person trying to turn an outcome from off to on (e.g. "paper not accepted" to "paper accepted"), who can intervene to turn a single variable on. In this scenario, the only interventions that can make a difference are on variables that are *currently off*. If your co-author has already blacklisted the bad reviewer, then intervening to blacklist him cannot help get the paper accepted. Hence, a measure of intervention effectiveness should be weighted towards variables that are more likely to be off.

The second reason why abnormal causes are weighed more heavily is captured by the $Prob(C = 1)$ term in the denominator of Definition 2. This reason is a counterpart to the first

[14] This exposition assumes that candidate interventions always involve turning a variable on, as opposed to off—i.e. introducing a factor which was absent. Of course, effects are sometimes produced by *removing* a factor — e.g. erasing a student's criminal record may help them get into college. This possibility is easily incorporated into our framework; "on" and "off" can be reinterpreted, not as the presence or absence of a variable, but as the relevant settings of the variable in the current context. We return to this issue in the discussion.

reason. Recall that in our model (and in typical human judgments), actual causes must have been necessary for the outcome. For a variable to be necessary, it had to be "on". A lit match could not have been necessary for the fire that occurred if there were no lit match. Hence, by always considering variables which were necessary, we end up, on average, biasing our actual cause judgments towards variables that tend to be present rather than absent. Dividing by $Prob(C = 1)$ in each particular case—i.e. focusing on abnormal variables—corrects for this bias.

Actual causes should have common outcomes Finally, the correction term weighs variables more heavily when the outcome is more common (via $\frac{Prob(E=1)}{Prob(E=0)}$). The reason for this weighting is that necessity judgments and general intervention effectiveness are restricted to different types of situations. Our approximation assumes that people are only making necessity judgments in situations where the outcome happens ($E = 1$). However, the effectiveness of an intervention is, in our definition, measured in situations where the outcome doesn't happen ($E = 0$).

The $\frac{Prob(E=1)}{Prob(E=0)}$ term corrects for this difference. It follows from Bayes' rule that, when conditioning observations on some criterion (e.g. $E = 1$), you divide by the probability of that criterion. In some sense, then, people's necessity judgments have been "divided" by $Prob(E = 1)$; multiplying by $Prob(E = 1)$ undoes that conditioning. Moreover, we want our measure of intervention effectiveness to be re-conditioned on $E = 0$; hence we divide the measure by $Prob(E = 0)$.

Importantly, the $\frac{Prob(E=1)}{Prob(E=0)}$ term provides a normative explanation for the supersession effect. Consider again the two-variable conjunctive case used in Icard et al. (2017) and Kominsky et al. (2015). An alarm goes off if and only if two employees arrive at 9 AM. The supersession effect is that people consider one employee (Billy) less causal if the other employee (Suzy) *rarely* arrives at 9 AM. This follows from our model: If Suzy rarely arrives at 9 AM, then, all else being equal, the alarm will go off less often—and hence Billy should be considered less causal.[15] We expand on this result in our experiment, and show that supersession takes a complex, non-linear form which is captured by our model.

Notably, this explanation of supersession differs from prior accounts. Supersession has been explained by assuming that, in addition to requiring causes to be necessary, people also require them to be "robustly sufficient"—i.e. sufficient across a variety of circumstances (Kominsky et al., 2015; Woodward, 2006). If Suzy usually arrives at 9 AM, it makes the alarm more likely to go off (the explanation considered here); but it also makes Billy more likely to be sufficient to make the alarm go off (the robust sufficiency explanation). Either of these theories, then, could explain the supersession effect in the two-variable conjunctive case.[16]

iii. Summary

We have argued that judging actual causes of past outcomes can be used to efficiently approximate the effectiveness of potential interventions, if those judgments follow three guidelines: Causes should be necessary, statistically rare, and linked to common outcomes. These guidelines offer a functional explanation for three prominent empirical patterns: necessity, abnormal selection, and supersession. Next, we test the fit of the SAMPLE model to human data across parametric variations of a very basic causal system, comparing its performance against current alternatives.

[15] Of course, this fact should make Suzy less causal also; but her loss in causal status from the $\frac{Prob(E=1)}{Prob(E=0)}$ term is outweighed by her gain from the $\frac{Prob(C=0)}{Prob(C=1)}$ term. We develop the precise predictions of our model in this case in our experiment.

[16] An attentive reader may wonder: What differs between our account of actual causation, and the robust sufficiency account? It seems like "robustly sufficient" is just another way of describing the general effectiveness of an intervention. The key difference is that, on our model, people don't explicitly employ a robust sufficiency criterion when making actual causal judgments. Rather, something akin to "sufficiency robustness" is the value which actual causal judgments are designed to approximate.

V. Experiment

The SAMPLE model predicts that, in order to approximate the effectiveness of potential interventions, actual causal judgments should take a particular form. To repeat our definition, the extent to which C is a cause of E in situation u_i should be:

$$AC_{SAMPLE}(C \rightarrow E, u_i) = \begin{cases} \frac{Prob(C=0)}{Prob(C=1)} \frac{Prob(E=1)}{Prob(E=0)} & \text{if } C \text{ was necessary for } E \\ 0 & \text{otherwise} \end{cases}$$

The importance of the necessity criterion has been previously established (Gerstenberg et al., 2014; Kahneman & Tversky, 1982; Wells & Gavanski, 1989). Here, we test the latter component of the definition: the correction term $\frac{Prob(C=0)}{Prob(C=1)} \frac{Prob(E=1)}{Prob(E=0)}$. To test whether actual causal judgments adhere to this form, we consider a simple causal system that isolates the influence of the correction term.

Consider a system where two variables combine conjunctively to produce an outcome (Figure 4A). This system is appealing because, in situations where the outcome is on, both variables are always necessary—thus, the predicted responses come entirely from the correction term in the SAMPLE model, and therefore provide a strong test of its form. As discussed above, prior research has shown that, in this causal structure, people judge a variable more causal when its occurrence was less likely (i.e. abnormal selection), and they judge it less causal when the other variable's occurrence was less likely (i.e. supersession). However, prior work has largely restricted itself to qualitative manipulations—e.g. comparing cases where a variable is rarely present to cases where it is almost always present.

An advantage of our approach is that it makes quantitative predictions. Indeed, the SAMPLE model predicts that people's causal judgments will be a complex, nonlinear function of the prior probabilities of the variables. We describe the SAMPLE model's predictions in detail below. To test whether people's causal judgments aligned with these predictions, we gave participants a vignette that included a causal system like that described above, and systematically manipulated the prior probability of the two potential causes. We asked participants to rate the extent to which they thought one of the variables was the cause of the outcome. We then correlated the predictions of the SAMPLE model (and the predictions of several alternative models) with people's responses. We find that people exhibit nonlinear variation in their responses, and that this variation is most closely tracked by the SAMPLE model.

i. Methods

We presented participants with the following vignette:

> A person, Joe, is playing a casino game where he reaches his hand into two boxes and blindly draws a ball from each box. He wins a dollar if and only if he gets a green ball from the left box and a blue ball from the right box.

(See Figure 4B; the entire vignette is reproduced in Figure 9 in the Appendix.)

We also presented images of the two boxes, showing the percentage of green balls in the left box and blue balls in the right box. By manipulating these images, we manipulated the prior probability that Joe draws a green ball or a blue ball. We will use `green` and `blue` to represent whether Joe drew a green ball and a blue ball, respectively. We divide the continuous space of probability into ten values: $Prob(\texttt{green} = 1) = 0.1$, $Prob(\texttt{green} = 1) = 0.2$, ... , $Prob(\texttt{green} = 1) = 1$. We divided the probability of `blue` similarly.[17]

For each observation of the system, we randomly selected a prior probability for `green` and `blue` (the two are independent), and showed the participant the corresponding image. We then tell the participant: Joe draws a green ball from the left box and a blue ball from

[17] We excluded zero because the outcome that participants actually observe is that Joe wins, which cannot happen if drawing either a green ball and a blue ball have probability zero.

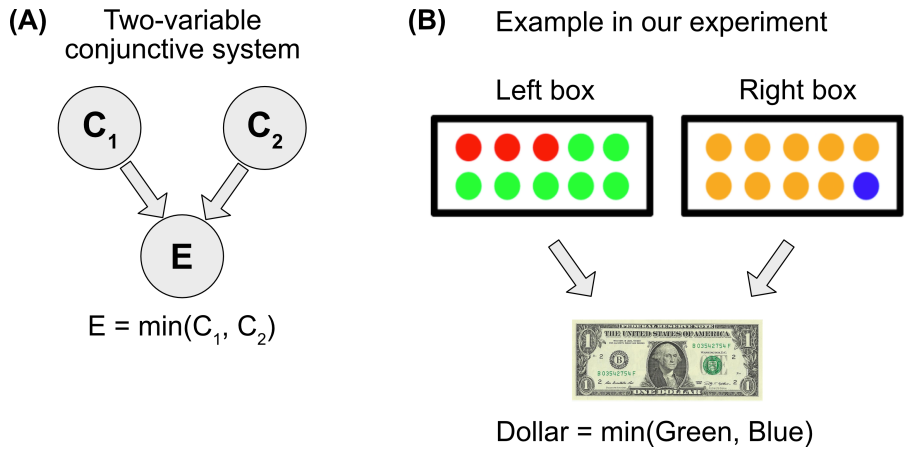


Figure 4: *(A) Causal structure with two variables that are individually necessary and jointly sufficient to produce an outcome. The outcome equals 1 if and only if both C_1 and C_2 equal 1. (B) The example system used in our experiment. Joe wins a dollar if and only if he draws a green ball from the left box and a blue ball from the right box.*

the right box, and wins a dollar. Finally, we asked the participant to rate the extent to which they agree with this statement: "Joe drawing a green ball from the left box caused him to win the dollar." (The participant is given a scale from 1 to 9, where 1 represents "totally disagree" and 9 represents "totally agree".) Thus, `green` was the "focal" variable whose causal status we asked people to judge, and `blue` was the "alternative" variable.

Participants All participants were recruited through Amazon Mechanical Turk. They gave informed consent, and the study was approved by Harvard's Committee on the Use of Human Subjects.

We employ a within-between subjects design. Each participant was given the vignette five times, each time with a randomly chosen probability setting. (We ensured that a participant did not see the same probability setting twice.) We ran $N = 999$ subjects, and collected a total of 4964 ratings. (No subjects were excluded, but some subjects did not complete all five ratings.) There are 100 different probability settings (10 for `green` crossed with 10 for `blue`), which gives an average of about 50 ratings per probability setting.

Analysis To compare people's responses with the model predictions, we first enumerate the predicted causality rating for `green` in each of the probability settings, according to our model and several alternative models. This is done in detail in the next section.

Then, for each model, we compute a Pearson correlation between the model's predictions for each probability setting and participants' averaged response in that setting. We compare the resulting correlations of each model.

ii. Model predictions

Our model Let `green` and `blue` represent whether Joe draws a green and blue ball; `dollar` represent whether Joe wins a dollar; and $AC_{SAMPLE}(\texttt{green} \rightarrow \texttt{dollar})$ represent the predicted causal ratings of `green` when Joe wins. To make the following equations more readable, we will use $Prob(\texttt{green})$ and $Prob(\neg\texttt{green})$ as shorthand for $Prob(\texttt{green} = 1)$ and $Prob(\texttt{green} = 0)$ — i.e. to indicate the probabilities of drawing and not drawing green, respectively. (We will do the same for `blue`.) According to the SAMPLE model, people's causal

judgments should take the following form:

$$AC_{SAMPLE}(\texttt{green} \rightarrow \texttt{dollar}) = \frac{Prob(\neg\texttt{green}) \cdot Prob(\texttt{blue})}{Prob(\neg\texttt{green}) \cdot Prob(\texttt{blue}) + Prob(\neg\texttt{blue})}$$

This formula follows from applying the definition above to the two-variable conjunctive causal structure. It has an intuitive interpretation: It is the probability that, given that Joe has not won a dollar, drawing a green ball would win him one. In other words, it is the general effectiveness of `green` for producing `dollar`—$GE(\texttt{green} \rightarrow \texttt{dollar})$.[18]

This causal structure, then, represents a special case for our model. In general, actual causal judgments are predicted to average to, not equal, the effectiveness of interventions (i.e. $GE(C \rightarrow E) = \frac{1}{k}\sum_{i=1}^{k} AC(C \rightarrow E, u_i)$). But when a variable (like `green`) is always necessary for an outcome, our model predicts that causal judgments will be identical in each episode—and hence equal to the average (in this case, $GE(\texttt{green} \rightarrow \texttt{dollar}) = AC_{SAMPLE}(\texttt{green} \rightarrow \texttt{dollar})$). In other words, by removing the influence of the necessity condition, we can test whether people's judgments align with the form predicted by the correction term.

The SAMPLE model's predictions are shown in Figure 5. It is helpful to unpack certain qualitative aspects of these predictions. First, as drawing a green ball becomes rarer, people should think drawing the green ball was more causal. This is abnormal selection. Second, as drawing a blue ball becomes more common, people should think drawing the green ball was more causal. This is supersession.

The novel predictions of the SAMPLE model concern the overall, nonlinear shape of the responses (Fig. 5). For instance, our model predicts relatively weak abnormal selection when the probability of drawing blue is close to (or equal to) 1—except for a steep jump from $Prob(\texttt{green}) = 1$ to $Prob(\texttt{green}) = .9$. Similarly, our model predicts relatively weak superseding when the probability of drawing green is close to 1—except for a steep jump from $Prob(\texttt{blue}) = .9$ to $Prob(\texttt{blue}) = 1$. These predictions are unique to the SAMPLE model, and follow from the model's functional rationale. As described above, in this simple causal structure, the SAMPLE model predicts that people's causal judgments should equal $GE(\texttt{green} \rightarrow \texttt{dollar})$—the probability that, given that Joe hasn't won, intervening to draw a green ball would make him win. When $Prob(\texttt{blue})$ is close to 1 (i.e. when Joe almost always draws a blue ball), this probability is consistently very high; if Joe hasn't won, it's probably because he didn't draw a green ball (and thus setting $\texttt{green} = 1$ would be a successful intervention). The only exception is when $Prob(\texttt{green})$ is also close to 1. Then, suddenly, it is no longer near-certain that `green` is the missing ingredient, and the probability of `green` starts mattering again. This explains why the SAMPLE model predicts relatively weak abnormal selection when $Prob(\texttt{blue})$ is close to 1, except for a steep jump when $Prob(\texttt{green})$ approaches 1 as well. (Similar logic also governs the supersession case.)

Normalized causal judgments According to our model, people internally compute a value that represents the extent to which `green` was an actual cause of `dollar`, which we represent with the function $AC_{SAMPLE}(\texttt{green} \rightarrow \texttt{dollar})$. So far, we've assumed that, when asked to rate the causality of `green`, people simply report this function's output. However, people may judge `green` in comparison to its alternative: `blue`. In other words, they may normalize the causal rating of `green` with respect to the other possible cause.

There are two reasons people might normalize their responses. One reason is pragmatic, or driven by particular features of the experimental design. The output of AC_{SAMPLE} has an unbounded upper range, but we ask people to rate causality on a bounded Likert scale. Moreover, in conversations involving two salient

[18] To see this, consider that there are two ways that the outcome could be off: Either `green` is off and `blue` is on, or `blue` is off. Only in the former case would setting `green` to 1 produce $\texttt{dollar} = 1$. AC_{SAMPLE} captures the probability of the former case.

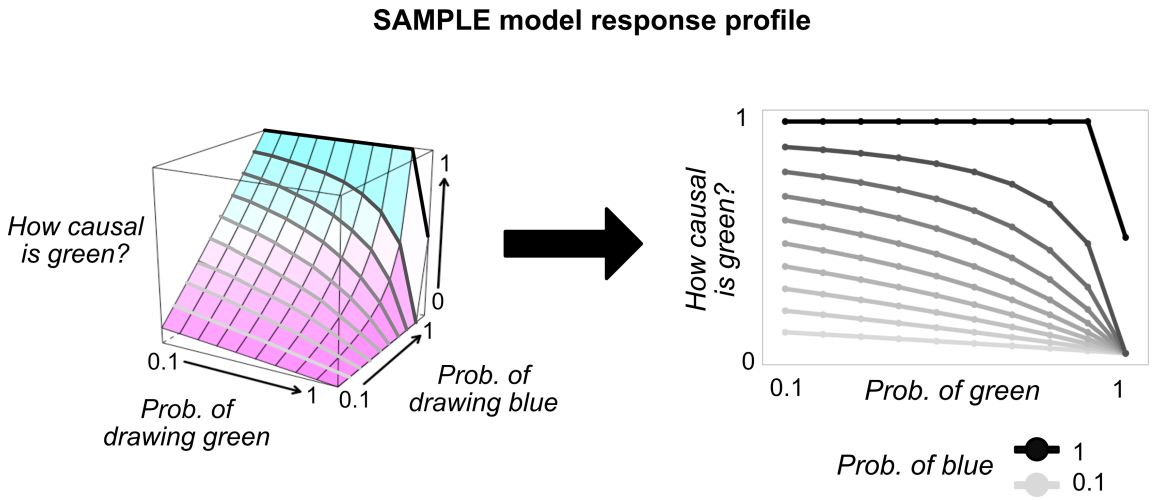


Figure 5: *The predicted response profile of the SAMPLE model, as a function of the prior probability that Joe would draw a green ball and a blue ball. The left graph shows the response profile in three dimensions; the right graph shows the same profile, collapsed into two dimensions. The SAMPLE model predicts abnormal selection (as the probability of* `green` *increases,* `green` *becomes less causal) and supersession (as the probability of* `blue` *increases,* `green` *becomes more causal); but it predicts that these effects will interact in nonlinear ways. (For ease of comparison, we will present the other model profiles — and the empirical results — in the two-dimensional format.)*

contributing variables, people may often assume that they are being asked to compare these variables to each other. It is thus possible that, although people employ the unnormalized output when privately interacting with the causal system, they often report normalized results when asked to explicitly judge a variable's causal status in a social setting.

But normalization may also serve a deeper purpose. As described above, a key function of estimating the general effectiveness of potential interventions would be to narrow down the set of interventions to consider at decision time. In other words, causality-driven estimates of $GE(C \rightarrow E)$ may help a decision-maker construct a "choice set" of plausible interventions available at decision time (Phillips & Cushman, 2017). Normalizing causal judgments would support this function. By comparing potential interventions to their alternatives, normalization can exaggerate the differences between them and highlight the top candidates to consider at decision time.

A natural way to accomplish normalization is to enter the value for $AC_{SAMPLE}(\texttt{green} \rightarrow \texttt{dollar})$ into a logistic (or "softmax") function:

$$AC_{SAMPLE-normalized}(\texttt{green} \rightarrow \texttt{dollar}) = \frac{e^{\beta \cdot AC_{SAMPLE}(\texttt{green} \rightarrow \texttt{dollar})}}{\sum_{color \in \{green, blue\}} e^{\beta \cdot AC_{SAMPLE}(\texttt{color} \rightarrow \texttt{dollar})}}$$

where β is a "gain" parameter that controls the strength of normalization. People's causal judgments would then reflect this normalized value. The idea that people normalize causal judgments is controversial (Gerstenberg, Goodman, et al., 2015; Kominsky et al., 2015; Lagnado et al., 2013). We do not claim it as a general rule. Nonetheless, the quantitative nature of our experiment allows us to test whether a normalized model better captures people's judgments in this causal structure.

The predictions of our normalized model are shown in Figure 6B. (To keep the model parameter-free, we fix β at 1.) The normalized model predicts an almost identical response profile as the non-normalized model, with two notable differences. The first is that the range is restricted; instead of rating the causality of

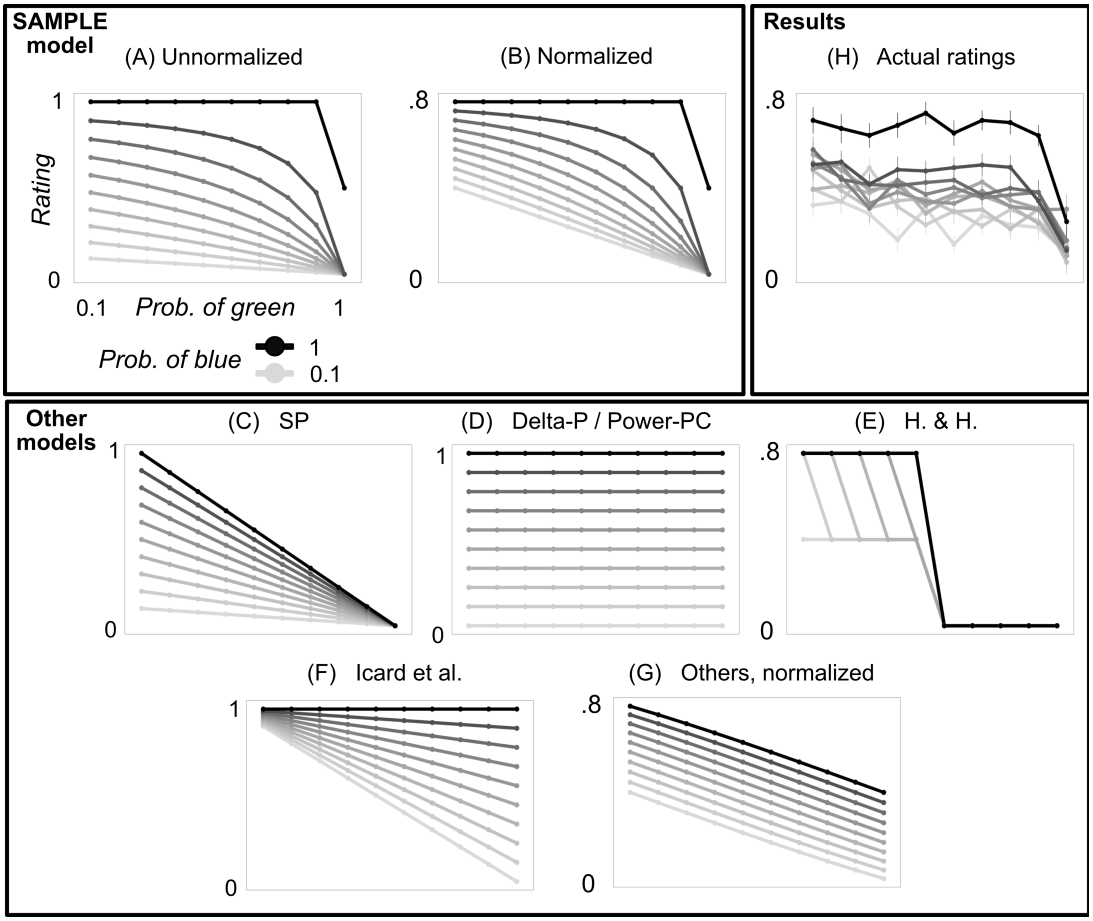


Figure 6: *The predicted response profiles of various models, as a function of the prior probability that Joe would draw a green ball and a blue ball. We employ a 2-dimensional representation here for ease of comparison across models. The x-axis depicts the probability of selecting a green ball; the line shade depicts the probability of selecting a blue ball; and the y-axis depicts the predicted (or empirical) causality rating for selecting the green ball. Axis scales are identical across graphs, with one exception: For the normalized models and empirical ratings, we restrict the y-axis to [0.2, 0.8] to get a clearer picture of the results. (A) The predictions of the SAMPLE model; (B) a normalized version of the SAMPLE model; (C) the SP model (Spellman, 1997); (D) the Delta-P (Jenkins & Ward, 1965) and Power-PC (Cheng, 1997) models; (E) a simplified version of the Halpern and Hitchcock (2015) model; (F) a simplified version of the Icard et al. (2017) model; and (G) a normalized version of the SP, Delta-P, Power-PC, and Icard et al. models (in this case, they are all identical). (H) People's average responses. There are roughly 50 observations per cell. (For a three-dimensional representation of the model predictions and empirical results, see Figure 10 in the Appendix.)*

green as zero when, e.g., $Prob(\texttt{green}) = 1$, the normalized model rates it at around 0.3. The second is that the normalized model predicts a more consistent abnormal selection effect, even when $Prob(\texttt{blue})$ is very low. In other words, the normalized model predicts that people will continue to rate drawing a green ball as more causal when it is rarer, even when blue balls are almost never drawn (see the downward slope on the light-gray lines in Figure 6B).[19] To preview our results, both of these distinctive features are observed in human data.

Alternative models Many models of actual causation do not consider abnormal selection and supersession effects, and predict that people's responses should not depend at all on the prior probabilities of green and blue. In other words, they predict a uniform surface. This includes the models of Hitchcock (2001), Halpern and Pearl (2005), and all process models (e.g. force-vector, or energy-transfer, models; Dowe, 2000; Walsh & Sloman, 2011; Wolff, 2007). We do not include these models in our analysis.

Some models do make quantitative predictions, but were designed explicitly for other causal structures, or have unknown parameters that make them difficult to analyze in this setting. Nonetheless, it is instructive to contrast them with our model. We include them here for exposition. (Though our model outperforms them in this setting, this result is not a critique of their applicability in the settings for which they were originally designed.)

There are three prominent covariation models that predict non-uniform surfaces in our experiment. The first, denoted SP, assigns causality to a variable to the extent to which observing the variable raises the probability of the outcome: $AC_{SP}(C \rightarrow E, u_i) = Prob(E \mid C) - Prob(E)$ (Spellman, 1997). In our case, this becomes: $AC_{SP}(\texttt{green} \rightarrow \texttt{dollar}) = Prob(\texttt{blue}) \cdot Prob(\neg\texttt{green})$ (Figure 6C).

The second is similar, but requires causes to raise the probabilities of their outcomes, relative to a state where the cause was absent. This model, denoted Delta-P, takes the form: $AC_{Delta-P}(C \rightarrow E, u_i) = Prob(E \mid C) - Prob(E \mid \neg C)$ (Jenkins & Ward, 1965; Perales & Shanks, 2007). In our case, this becomes: $AC_{Delta-P}(\texttt{green} \rightarrow \texttt{dollar}) = Prob(\texttt{blue})$ (Figure 6D).

The third builds on the second, but normalizes the rating by the probability that the event is absent when the cause is absent. This model, called Power-PC, takes the form: $AC_{PPC}(C \rightarrow E, u_i) = \frac{Prob(E|C) - Prob(E|\neg C)}{Prob(\neg E|\neg C)}$ (Cheng, 1997). In our case, this is identical to Delta-P: $AC_{PPC}(\texttt{green} \rightarrow \texttt{dollar}) = Prob(\texttt{blue})$ (Figure 6D).

Again, these models were designed for different causal structures (Griffiths & Tenenbaum, 2005), and are most naturally applied to type causation, not the token causes we consider here. Nonetheless, the ways in which they differ from our model serve as a useful and appropriate comparison.

There are two additional models that are also helpful as comparisons. The first model comes from Halpern and Hitchcock (2015). Halpern and Hitchcock argue that when people evaluate counterfactuals and consider alternative possible worlds, they prefer to imagine worlds that are more normal, where unlikely things haven't occurred (Kahneman & Miller, 1986; Kahneman & Tversky, 1982). For instance, people more tend to imagine a world in which an arsonist didn't light a match than a world in which there were no oxygen in the air. Or, in our case, if it's rare to draw a green ball but common to draw a blue ball, they would tend to imagine an alternative world in which a green ball was not drawn.

Halpern and Hitchcock (2015) argue that this observation can explain the dependence of actual causation on the prior probabilities of the variables. To determine whether a variable C

[19] Our model, in both its unnormalized and normalized forms, does not technically assign causal ratings in the situation where it is certain that the agent will draw both a green and blue ball—i.e. $Prob(\texttt{green}) = Prob(\texttt{blue}) = 1$. (In this case, the denominator in $AC_{our-model}(\texttt{green}, \texttt{dollar})$ is zero.) However, if people do not assign a probability of exactly 1—e.g. if they assign a probability of 0.999—to $Prob(\texttt{green})$ and $Prob(\texttt{blue})$, then the causal rating of green approaches 0.5. Thus, we assign a rating of 0.5 in this case. The results change little if this probability setting is dropped from the analysis.

was counterfactually necessary for an outcome, we consider an alternate world in which the variable is off. Call this world w_C. Halpern and Hitchcock propose that, for a particular observation of a causal model, people will only call a variable an actual cause if it is counterfactually necessary for the outcome[20], and if the possible world considered in determining necessity—w_C—was more normal than the actual world. Moreover, if two variables C_1, C_2 meet these criteria, then C_1 will be considered more of an actual cause if w_{C_1} is more normal than w_{C_2}.

From Halpern and Hitchcock's model, we can derive a predicted response profile in our experiment (Figure 6E). If the probability of drawing a green ball is greater than 0.5, then the green ball should never be considered an actual cause, because $w_{\texttt{green}}$–the world in which green is off–is less likely than the actual world. (In the actual world, a green ball was drawn; in $w_{\texttt{green}}$, a green ball was not drawn. If $Prob(\texttt{green} = 1) > .5$, then, all else being equal, $w_{\texttt{green}}$ is less likely than the actual world.) This result is depicted in the lowest region of Figure 6E. If the probability of drawing green is less than 0.5 and also less than the probability of drawing blue, then green is assigned a high causal rating, because $w_{\texttt{green}}$ is more likely than $w_{\texttt{blue}}$. This is depicted in the highest region of Figure 6E. Finally, if the probability of drawing green is less than 0.5 but greater than the probability of drawing blue, then green is causal but less causal than blue; it is assigned an intermediate rating.

Formally, on one plausible interpretation of Halpern & Hitchcock's model, the degree to which green = 1 was the actual cause of Joe's win is:

$$AC_{HH}(\texttt{green} \rightarrow \texttt{dollar}) = \begin{cases} 0 \; if \; Prob(\texttt{green} = 1) > .5 \\ x_{high} \; if \; Prob(\texttt{green}) < Prob(\texttt{blue}) \leq .5 \\ x_{mid} \; if \; Prob(\texttt{blue}) < Prob(\texttt{green}) \leq .5 \end{cases}$$

where $0 < x_{mid} < x_{high} \leq 1$.

Halpern & Hitchcock's model is ordinal, and so it cannot assign precise numbers for x_{high} and x_{mid}. Here, for simplicity, we assume that $x_{high} = 1$ and $x_{mid} = \frac{1}{2}$. (Fitting these values to the empirical results does not substantially improve the model fit.)

Another model we consider comes from Icard et al. (2017). Icard et al. offer another variation of the counterfactual approach. They propose that, when evaluating actual causality, people sample a counterfactual world to consider with probability proportional to how likely that world is. For instance, to evaluate whether green = 1 was the cause of Joe's win, they sample a world in which green = 0 (he didn't draw a green ball) in proportion to $Prob(\neg\texttt{green})$, and a world in which green (he did draw a green ball) in proportion to $Prob(\texttt{green})$. Then, crucially, people evaluate a different counterfactual depending on what they sampled. If they sampled a world in which green = 0, they evaluate whether green was necessary (holding all else about the actual world fixed); if they sampled a world in which green = 1, they evaluate whether green is, in general, sufficient (allowing other variables to vary).

According to Icard et al., while the probability of sampling a counterfactual world with green = 0 is monotonically related to $Prob(\texttt{green} = 0)$, the two are not necessarily equal; there may be other factors that influence which worlds come to mind. This feature makes it difficult to derive quantitative predictions in our experiment. Nonetheless, if we make the improbable but simplifying assumption that the probability of sampling counterfactual worlds comes directly from the prior probability of green, then we can derive a re-

[20] Halpern and Hitchcock's model is more intricate than this. Building on Halpern and Pearl (2005), it says that a variable must be necessary after making some allowable modifications to the situation. We return to these intricacies in the discussion, when considering cases of overdetermination.

sponse profile to compare to our model:

$$AC_{Icard}(\texttt{green} \rightarrow \texttt{dollar}) = Prob(\neg\texttt{green}) \cdot Prob(\texttt{green} \text{ was necessary}) + Prob(\texttt{green}) \cdot Prob(\texttt{green} \text{ would be sufficient})$$

which reduces to:

$$AC_{Icard}(\texttt{green} \rightarrow \texttt{dollar}) = 1 - Prob(\texttt{green}) \cdot Prob(\neg\texttt{blue})$$

The predictions of this simplified version of Icard et al.'s model are shown in Figure 6F. Icard et al.'s model predicts many qualitatively similar effects as our model, but the overall shape is different. Most obviously, their model predicts strong supersession when $Prob(\texttt{green}) = 1$, while ours predicts very little.

Finally, for completion, we also include a normalized version of three of the alternate models: SP, Delta-P / Power-PC, and Icard et al. (We leave out the normalized version of the Halpern & Hitchcock model, because it is very similar to the unnormalized version.) We normalize these alternate models in the same way we normalized ours: by entering their AC measures into a logistic function. Interestingly, in this case, the normalized versions of all three of these alternate models are identical to each other (the causal ratings for green are all $\frac{e^{Prob(\texttt{blue})}}{e^{Prob(\texttt{blue})}+e^{Prob(\texttt{green})}}$). Their predicted response profile is shown in Figure 6G.

iii. Results

People's average ratings are shown in Figure 6H, scaled from their original range of 1 to 9 to a range of 0 to 1. People clearly exhibit abnormal selection; on average, they rate drawing green as more causal when drawing green was less likely. They also clearly exhibit supersession; on average, they rate drawing green as more causal when drawing blue was more likely.

To demonstrate both of these effects statistically, we estimated a linear mixed effects model, regressing the causal ratings of green on the prior probabilities of drawing green and blue, with random intercepts and slopes for each subject. On average, people indeed rated green as more causal when $Prob(\texttt{green})$ was lower ($\beta = -0.02, SE = 0.002, t(959.9) = -11.8, p < .001$) and when $Prob(\texttt{blue})$ was higher ($\beta = 0.03, SE = 0.001, t(949.5) = 18.9, p < .001$).

More interesting, however, are the nonlinear patterns. As predicted by our model, people exhibit markedly reduced abnormal selection and superseding when $Prob(\texttt{blue}) = 1$ and $Prob(\texttt{green}) = 1$, respectively. The one exception, as predicted by our model, is when the probabilities of drawing green and blue are both close to 1. In that case, people show steep abnormal selection and supersession.

Notably, when comparing the normalized and unnormalized versions of our model, people's ratings were most consistent with the normalized version. The absolute value of their ratings largely fell within the normalized model's restricted range (around 0.3 to 0.7), and they exhibited consistent abnormal selection, even when drawing blue balls was very rare. (We confirmed the latter observation by running a linear regression on just the observations where $Prob(\texttt{blue})$ was at its lowest probability. As predicted by the normalized, but not unnormalized, model, people still rated drawing a green ball as more causal when it was rarer, $\beta = -0.02, t = -3.28, p = .0014$.) These results provide some evidence that, in this experiment, people are likely normalizing their responses. An interesting question for future research is when, and why, people normalize when making judgments of actual causation.

Item-level correlations between model and data To compare people's responses to the model predictions, we computed the correlation between the empirical ratings and each model's predictions, across settings of $Prob(\texttt{green})$ and $Prob(\texttt{blue})$ (Figure 7). Our unnormalized model had a correlation of 0.86 ($t(98) = 16.6, p < .001$), and our normalized model had a correlation of 0.87 ($t(98) = 17.7, p < .001$). Both correlations were significantly higher than that of the next-best model

(the normalized alternate models; Williams' t-test, ts > 2.5, ps $< .01$), and they did not significantly differ from each other ($t < 1$, $p = 0.42$).

VI. Discussion

What are actual causation judgments for? We argue that a key role these judgments serve is to accumulate evidence about generally effective interventions in a computationally efficient manner. Consistent with this possibility, our SAMPLE model captures participants' quantitative causal judgments in a simple causal structure. It also provides a natural, functional account of three qualitative patterns widely observed in judgments of actual causation: necessity, abnormal selection, and supersession. This match between the predictions of a normative model of decision-making and the established patterns in judgments of actual causation suggest that they share a common functional design.

More generally, our model addresses a basic question about the nature of causal judgment. As noted by Hitchcock (2017), an organism's fate is determined by the future consequences of its actions. Decision-making must, in this sense, be "forward-looking". Yet humans appear obsessed with inherently backward-looking ascriptions of causal responsibility. Why?

The SAMPLE model bridges this gap by invoking the notion of a sampling approximation. Past observations of a causal system can act as samples which, when combined appropriately, enable a person to approximate the future effectiveness of potential interventions. Viewing actual causation as an implementation of this sampling approximation explains why a forward-looking enterprise like decision-making would be served by a backward-looking notion like actual causation (Lombrozo, 2006).

Our remaining discussion begins with a review of the assumptions made in our analysis, considering how these assumptions might be relaxed in extensions of the basic SAMPLE approach. Then, we describe the connections between the SAMPLE model and previous work, and highlight how our model may be more deeply integrated with previous approaches. Finally, we consider future directions and open questions.

i. Assumptions, extensions, and limitations

We have stated the SAMPLE approach in its simplest form, and tested only its most fundamental predictions. Here, we address several important simplifying assumptions, and point to how the model may be extended to overcome them. Some assumptions are relatively benign; others represent deeper commitments of the model, and may restrict the types of scenarios to which our analysis applies.

The nature of the approximation algorithm We made several assumptions about the nature of the approximation algorithm that can be easily relaxed. First, we proposed that people estimate the effectiveness of interventions by averaging actual causal judgments from all observed episodes (i.e. $GE(C \rightarrow E) \approx \frac{1}{k}\sum_{i=1}^{k} AC_{SAMPLE}(C \rightarrow E, u_i)$). Interpreted literally, this imposes a heavy memory burden, requiring people to remember all their past judgments. A more likely possibility is that people would maintain a "running average" of $AC_{SAMPLE}(C \rightarrow E, u_i)$, updating it as new episodes are observed.

Second, we proposed that people would weigh all past episodes equally when averaging their actual causal judgments together. If the effectiveness of interventions does not change over time, this uniform weighting is appropriate. But in real life, the effectiveness of interventions is likely a moving target—e.g. writing an engaging essay may be more important for getting into college now than it was fifty years ago. A simple, and commonly used approach in many cognitive models (e.g., prediction-error learning) is to weigh recent episodes more heavily, for instance by using a constant update rate (Daw, Niv, & Dayan,

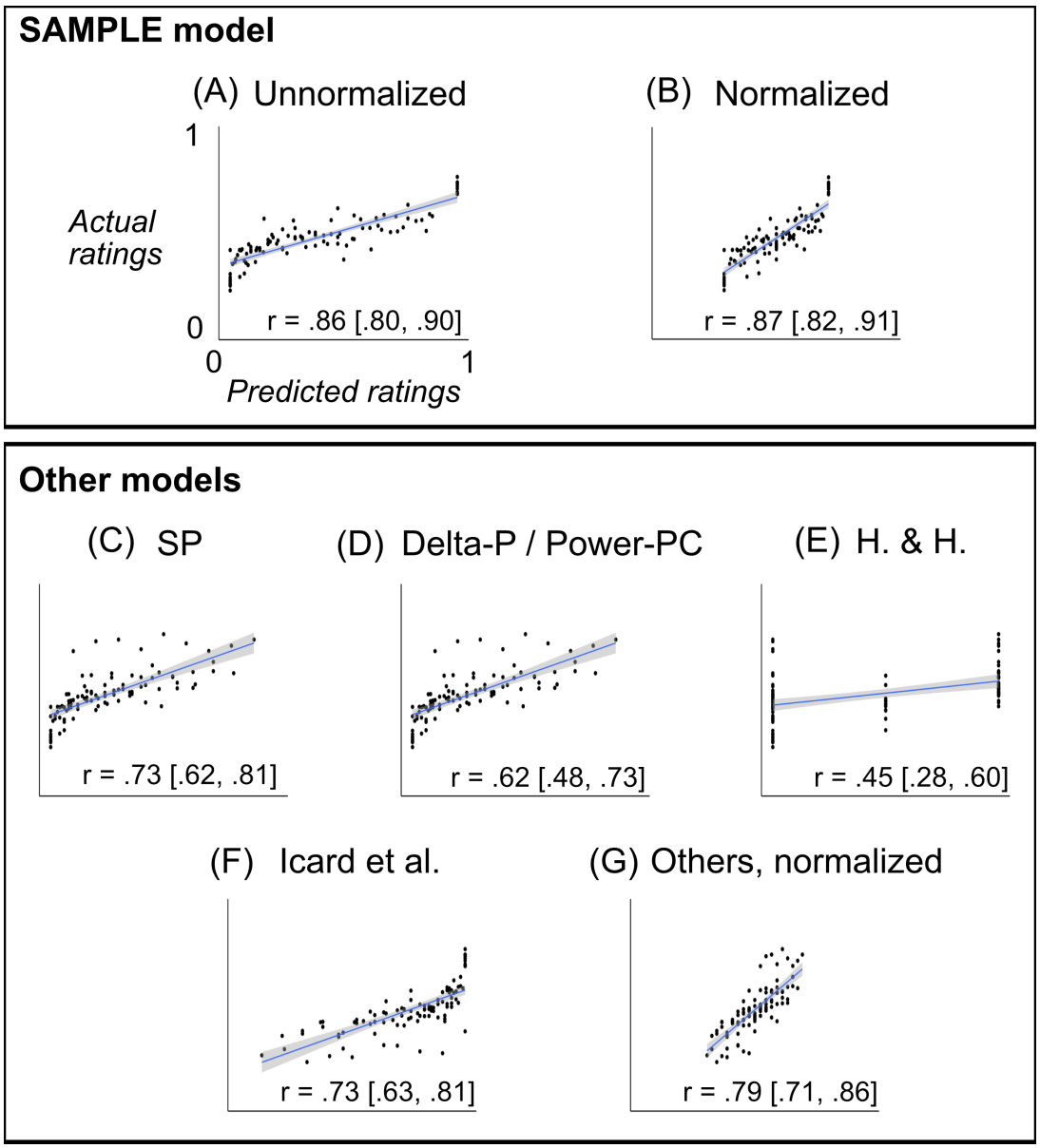


Figure 7: *Relationship between model predictions and empirical ratings. Brackets show 95% confidence intervals for Pearson correlation coefficients.*

2005).[21]

Third, we assumed that, when people judge causation in past episodes, they know the settings of all variables in the causal system. This unrealistic assumption is easily relaxed. Our algorithm only requires that people observe enough of the causal system to assess whether potential causes were necessary for the outcome. This is not typically an onerous requirement; as discussed above, people appear able to assess necessity in a wide variety of situations. (Moreover, even if a person only observed enough variable settings to assess the *probability* that a potential cause was necessary, that value would be sufficient for our approximation; in our definition of actual causation [Def. 2], the indicator function I(C was necessary for E) can be replaced with the weighting $Prob$(C was necessary for E).)

The nature of the decision problem We also made several assumptions about the nature of the decision problem that actual causation is functioning to solve. Some of these are less easily relaxed, and may set boundaries on the problem space in which actual causal judgments of the form we describe have value for decision-making.

Binary variables First, we assumed that all variables—causes and effects—were binary. This assumption is less restrictive than it sounds. All that is required is that the decision-maker *treat* the variables as binary in the relevant decision context. For instance, consider a categorical outcome with more than two categories, like a letter grade on a test. The variable `grade` has five possible values A, B, C, D, F, and yet it can still be treated as binary if the decision-maker's goal is to get an A on the test. In that case, the relevant outcomes are $\texttt{grade} = A$ and $\texttt{grade} \neq A$. (It's not necessary for the goal to be a single outcome setting; if the decision-maker's goal was to get an A or B, the relevant outcomes would be $\texttt{grade} \in \{A, B\}$ and $\texttt{grade} \notin \{A, B\}$.) Similarly, a continuous variable — say, the test's numerical grade — can also be treated as binary if the decision-maker's goal is to get above (or below) a threshold. In that case, the relevant outcomes are $\texttt{grade} > \textit{threshold}$ and $\texttt{grade} \leq \textit{threshold}$. (The same considerations apply to categorical or continuous causes; so long as the decision-maker conceptualizes potential interventions as binary, our analysis remains unchanged.)

Difficulties only arise when decision-makers genuinely conceptualize variables as continuous in the relevant decision context. For instance, a decision-maker's goal may be to maximize the numerical grade they receive on a test. In that case, our definition of an effective intervention is inadequate. Instead of maximizing the probability of the outcome's occurrence (i.e. $Prob(E_{C=1} = 1 \mid E = 0)$, a decision-maker must maximize the expected value of the outcome ($EV[E_{C=1}]$), perhaps conditional on the outcome currently occupying an undesired state. Although possible that the principles described in this paper will apply to such cases, this question is currently unresolved.

Omission Second, we assumed that, when judging potential causes or interventions, people only consider factors that were present, not absent. For instance, we assumed that the admissions officer would judge whether the presence of an engaging essay was a cause of victory, but not whether the absence of a criminal record was a cause. (Formally, a variable C was only considered a cause if $C = 1$, not $C = 0$, was necessary for the outcome.) This assumption allowed us to avoid tricky issues surrounding absent variables. But, in real life, people sometimes consider absences to be causal (Henne, Pinillos, & De Brigard, 2017; Stephan, Willemsen, & Gerstenberg, 2017; Willemsen, 2016; Wolff, Barbey, & Hausknecht, 2010). Moreover, a similar issue arises with absent *outcomes*. We assumed that people would only determine the causes of an outcome's presence (i.e. cases where $E = 1$); but people can

[21] Formally, let $\overline{AC}_{SAMPLE}(C \rightarrow E)$ indicate the running average. If, after observing u_{new} people are updating according to $\overline{AC}_{SAMPLE}(C \rightarrow E) \leftarrow (1 - \alpha)\overline{AC}_{SAMPLE}(C \rightarrow E) + \alpha AC_{SAMPLE}(C \rightarrow E, u_{new})$, then a constant α will ensure an exponential recency-weighting.

also consider the causes of an outcome's absence (where $E = 0$).

A straightforward extension of our framework accommodates many such cases. To repeat the basic result we've shown: Judging whether the presence of a variable caused the presence of an outcome helps approximate the probability that introducing the variable would produce the outcome. (Formally, judging $AC_{SAMPLE}(C = 1 \rightarrow E = 1)$ helps approximate $GE(C = 1 \rightarrow E = 1)$.) Similar logic can be applied to the cases of absent causes or absent outcomes. Judging whether the absence of an event caused an outcome helps approximate the probability that removing the event would produce the outcome. (Formally, judging $AC_{SAMPLE}(C = 0 \rightarrow E = 1)$ helps approximate $GE(C = 0 \rightarrow E = 1)$.) For instance, by judging whether the absence of a criminal record caused past applicants to get accepted, the admissions officer can estimate the effectiveness of ensuring that her children stay out of legal trouble (i.e. intervening to set `CriminalRecord` = 0). Similarly, judging whether a variable caused the absence of an outcome helps approximate the probability that introducing that variable would remove the outcome.

The only difference in these cases arises in the correction term. When judging whether the absence of a variable caused an outcome—$AC_{SAMPLE}(C = 0 \rightarrow E = 1)$—the $\frac{Prob(C=0)}{Prob(C=1)}$ term is replaced by $\frac{Prob(C=1)}{Prob(C=0)}$. In other words, when judging whether the absence of a factor caused an outcome, our model predicts that people should consider *common* factors more causal. This is intuitive—if a student is the only one without a criminal record, that fact will likely be considered more causal. (Another way to put it is that people should consider the absence more causal when it is abnormal—i.e. when the variable itself is common. In this way, the abnormal selection effect is preserved.) Similarly, when judging whether a variable caused the absence of an outcome—$AC_{SAMPLE}(C = 1 \rightarrow E = 0)$—the $\frac{Prob(E=1)}{Prob(E=0)}$ is replaced by $\frac{Prob(E=0)}{Prob(E=1)}$. In other words, when determining the causes of absent outcomes, our model predicts that the supersession effect should flip; in conjunctive scenarios, people should consider a variable more causal when the alternative is rarer. To our knowledge, this has not been tested.

An upshot of these considerations is that different types of causal judgments—causing an outcome versus preventing an outcome, for instance—may function to approximate the effectiveness of different types of interventions. By judging whether well-written essays caused past applicants to get accepted, you can discern whether improving your writing will help you get accepted; by judging whether a lack of extracurriculars prevented past applicants from getting accepted, you can discern whether quitting the debate team will hurt your chances. Of course, the picture is more complex than this. Sometimes the type of causal judgment is mismatched with the desired intervention; e.g. sometimes people judge which variables caused the presence of an outcome, even though the outcome is one they want to prevent. (A liberal pundit might angrily describe what caused a conservative political candidate's victory.) We leave it to future work to unravel the connections between types of causal judgments and types of interventions.

Ease of assessing necessity A third assumption of our approach is that people estimate intervention effectiveness in scenarios where necessity is, on average, easier to identify and assess than sufficiency. In other words, we assumed that it would be easier to identify variables which were necessary (or assess whether a potential cause was necessary) for an outcome than to identify variables which were sufficient for the outcome. This assumption helped motivate why people would use necessity judgments, rather than sufficiency judgments, to estimate the effectiveness of interventions—and hence why actual causal judgments would take the form that they do (Gerstenberg et al., 2014; Halpern & Pearl, 2005).

This assumption is clearly met in strictly con-

junctive situations, where all variables are necessary for an outcome (e.g. a fire occurs just in case there's oxygen, and dry wood, and a lit match, and so on). But our analysis is not restricted to purely conjunctive causal structures. There are many causal structures with disjunctive elements—"OR" parameterizations—in which necessity is still easier to identify and assess. Consider again the example of shooting a basketball. There is more than one way to make a successful shot—you can release the ball lower or higher; you can jump towards the basket, or away from it; etc.—suggesting that there are disjunctive elements in this causal system. Yet, it is remarkably easy to identify necessary variables (and to assess whether a particular variable was necessary), and remarkably difficult to identify sufficient ones. This pattern is likely replicated in any real-world structure where there are few sets of jointly sufficient variables, each of which is composed of many individually necessary variables (Mackie, 1974).

Of course, there are also many causal systems where this assumption does not hold; that is, where sufficiency is easier to compute than necessity. In such cases, people may shift to sufficiency judgments to estimate intervention effectiveness, and our model may be less applicable. But even then, the necessity judgments described here may still function as a supplementary source of information about intervention effectiveness. Necessity-based actual causation judgments need not be the *only* method of approximating intervention effectiveness; our claim claim is that actual causation is one important cognitive tool that fills that role.

No-confounding assumption Finally, we assumed that actual causation primarily functions in causal structures that obey the "no-confounding" assumption: the effect on desired outcome E from intervening on potential cause C does not depend on the current setting of C. Graphically, this assumption is guaranteed to hold if the cause and effect have no common ancestor; there are no "confounds" connecting them. This assumption is an indelible commitment of the SAMPLE approach to actual causation; without it, necessity-based actual causation judgments cannot accumulate evidence about the effectiveness of interventions.

Fortunately, it is reasonable to suppose that actual causal judgments are at least partially predicated on the no-confounding assumption. The no-confounding assumption is met (or partially met) often enough to warrant its widespread adoption in fields like epidemiology, law, and econometrics (Pearl, 1999, 2000). Moreover, from a psychological perspective, the no-confounding assumption is weaker than the assumptions made in popular models like Power-PC (Cheng, 1997; Luhmann & Ahn, 2005), and people are anecdotally hesitant to make causal judgments when it is violated (Luhmann & Ahn, 2005).[22] Finally, from a functional perspective, adopting the no-confounding assumption is the only way to extract information about intervention effectiveness from necessity judgments. (Tian and Pearl (2000) show that, without the no-confounding assumption, the probability that an intervention successfully produces an outcome is not bounded by information about necessity.) If people face many situations where they (a) must choose interventions to produce outcomes, and (b) can identify necessary variables from past experience, then a reasonable approach is to adopt the no-confounding assumption and extract whatever information can be gleaned from that past experience.[23]

[22] An important caveat to these observations is that the no-confounding assumption is most often adopted in treatments of *type* causation, not actual causation. The extent to which the assumption translates to cases of actual causation has been less explored.

[23] Note that if the no-confounding assumption fails, necessity judgments will not systematically bias people *away* from effective interventions; they will simply be uninformative, and add noise to the estimate. Hence, a cost-benefit analysis would reasonably conclude that making the no-confounding assumption, and using necessity judgments, is worthwhile.

ii. Relationship to prior work

Our approach interfaces with prior work in numerous ways. Here, we review those connections, and describe how the SAMPLE model builds on and integrates with other approaches.

Prior work on normality Prior treatments of actual causation propose that it is closely related to the problem of finding good interventions. Hitchcock and Knobe (2009) suggested that the two concepts are linked, and argued that this linkage could explain the effects of normality on causal judgment (see also Woodward, 2003). This idea is at the core of our proposal. Our model elaborates and formalizes this idea, showing that an actual causation algorithm designed to identify good interventions would exhibit many observed effects of normality, as well as other empirical patterns.

Our model focuses exclusively on the effects of statistical abnormality on causal judgment—i.e. the fact that people are more likely to select rare events as causes in conjunctive settings. But the effects of normality go beyond statistics: People also believe that prescriptively abnormal events are more causal (Hitchcock & Knobe, 2009; Icard et al., 2017; Phillips, Luguri, & Knobe, 2015). Agents who violate prescriptive norms, or objects which violate functional norms, are considered more of a cause of the resulting outcomes (Hitchcock & Knobe, 2009; Phillips & Kominsky, 2017). Prescriptive norms also apply to the supersession effect; an agent in a conjunctive setting is considered less causal if its *counterpart* violated a prescriptive norm (Kominsky et al., 2015). Wherever statistical norms have an effect, prescriptive norms seem to have a similar effect (Bear & Knobe, 2017)

How can our model account for these effects? As noted above, our model is primarily a functional account of actual causation; it is likely missing elements of the mechanism underlying real causal judgments. One salient element comes from Bear and Knobe (2017), who argue that, across many cognitive tasks, people replace estimates of a variable's base rate—i.e. $Prob(C = 1)$—with an estimate of how "normal' the variable is, where "normal' integrates both statistical and prescriptive norms. If this is true, then prescriptive normality would have a similar effect as statistical normality in our model (Icard et al., 2017). Of course, it is also possible that there is some specific functional reason that actual causal judgments would incorporate prescriptive normality (Hitchcock & Knobe, 2009); we leave it to future work to investigate this possibility.

Prior work on actual necessity As described above, much prior research shows that causal judgment depends on an analysis of necessity. There is, however, a particular feature of necessity judgments for which our model offers a novel justification. Even in cases in which the circumstances that made a variable necessary were highly engineered or unusual, or where a small change would have rendered the variable unnecessary, people are still willing to attribute causation, as long as the value of the variable was necessary in the actual circumstances that occurred.

For instance, imagine a school which gives an annual award either to students who get an A in English, or students who get an A in math. Almost always, the school gives the award for math; but this year, the school decides to give it for English. Lilly, a student at this school, gets A's in both English and math, so she wins the award. It seems clear that Lilly's English grade, not her math grade, was the cause of her winning the award — even though the circumstances which rendered her English grade necessary were highly peculiar. This feature of causal judgment is known as "actual necessity": A cause need only be necessary in the *actual* situation in which it occurred, even if the features of the situation that rendered it necessary were highly unusual (Icard et al., 2017; Spellman & Kincannon, 2001; Woodward, 2006).

This feature follows naturally in the SAMPLE model because actual causation judgments are implementing a sampling approximation.

The proportion of instances in which a putative cause was necessary historically is used as a guide to estimate the proportion of time that it will be an effective intervention in the future (modulo the appropriate correction term). If a variable is not often necessary in general, then of course one will not often encounter specific situations in which it is necessary. Indeed, the very approach of employing sampling approximation would actually be undermined by attempting to "correct" specific judgments of past necessity in light of their general necessity; after all, specific cases arise just in proportion to their general propensity. Moreover, the sampling approximation is cognitively efficient precisely because it avoids the computational demands of deriving general statistics about necessity, intervention effectiveness, or anything else. The empirical phenomenon of "actual necessity" thus follows directly from the present approach to causal judgment.

Overdetermination and preemption A major shortcoming of the SAMPLE model is its inadequacy in cases of overdetermination or preemption. As discussed above, these are cases in which two events are each sufficient to produce an outcome, and both happen (or one happens, and the other would have happened had the first failed). In these cases neither event was necessary for the outcome; hence, if people are implementing the computational approach described here, they should consider neither event a cause. But in fact, people think they are causal (Halpern, 2016; Spellman & Kincannon, 2001).

Many accounts have been developed to explain people's judgments in cases of overdetermination and preemption (Gerstenberg, Goodman, et al., 2015; Halpern, 2016; Halpern & Pearl, 2005; Lewis, 2000; Stephan & Waldmann, 2016). One possibility is that, even though our model accurately describes one important aspect of the functional design of actual causal judgments, causal judgment in overdetermined cases is explained by a different mechanism with a different functional purpose.

One explanation, for instance, comes from process theories. These assign causality to events that were connected to the outcome via some physical process (Dowe, 2000; Wolff, 2007; Wolff et al., 2010). According to process theories, if two soldiers simultaneously shoot a prisoner in the head, both events were physically connected to the prisoner's death—and hence both should rightly be considered a cause. Process theories thus offer a natural explanation for people's intuitions in cases of over-determination. The operation of process-based mechanism could be reconciled with our account if, as is often argued, people possess two different concepts of causation: one dependence concept and another process concept (Hall, 2004; Lombrozo, 2010). On this view, our model would offer a functional account of people's dependence concept, which is simply overshadowed by the process concept in cases of overdetermination/preemption. Despite their appeal, however, process theories have been criticized on various grounds (e.g., difficulty in dealing with cases of double prevention, Gerstenberg et al. 2014; Schaffer 2000, 2004).

Alternatively, it is possible that the basic function of the SAMPLE model is preserved in overdetermination cases with appropriate modifications to the simple algorithm described here. In other words, with some additional computation, people may be able to extend the sample-based method of estimating the effectiveness of interventions even from cases where no individual variable was necessary to produce an effect. We briefly sketch two possible approaches.

First, people might not only be able to assess whether any given variable setting was necessary to produce an effect; in some cases, they might invest additional effort to assess whether any given variable was sufficient, given certain background conditions (Icard et al., 2017; Woodward, 2006). As noted above, necessity judgments are an efficient way to approximate the effectiveness of interventions, but those approximations can be improved by also considering sufficiency. When two soldiers simultaneously shoot a prisoner in the head, both are

sufficient for the prisoner's death (holding constant certain background conditions). If people are making additional sufficiency judgments of this form, they might therefore judge the soldiers causal (Spellman & Kincannon, 2001). Potentially, an appropriate correction term applied to such judgments would allow them to accumulate information about general intervention effectiveness.

A related possibility is that people are not limited to judging whether a variable setting was necessary in the actual circumstances in which it occurred; in some cases, they may also judge whether it was necessary in slightly modified circumstances. For instance, if two soldiers simultaneously shoot a prisoner in the head, neither was necessary in the actual circumstances—but the first soldier would have been necessary had the second soldier not pulled the trigger (Halpern & Hitchcock, 2011; Halpern & Pearl, 2005; Hitchcock, 2001; Lucas & Kemp, 2015).[24] Exploiting this fact, people could (*i*) anchor on the actual state of affairs, (*ii*) adjust minimally to obtain a counterfactual state of affairs in which some variable setting would indeed have been necessary, and then (*iii*) apply our algorithm in the minimally counterfactual state of affairs. This would require the application of an additional correction term that adjusts for the relative frequency of the actual versus the counterfactual states of affairs. More concretely, by considering the counterfactual state in which the second soldier didn't shoot, people can sharpen their estimates of intervention effectiveness without having to wait to observe such an episode. Instead, they generate a simulated "observation" by adjusting minimally from the state of affairs they actually observed.[25]

Both of these approaches—the sufficiency analysis and the counterfactual necessity analysis—serve the same basic function as the SAMPLE model. By using a generative model of the situation to infer what would have happened in plausible altered circumstances, people may be generating additional "observations" of the causal system (Blanchard, Vasilyeva, & Lombrozo, 2017; Vasilyeva et al., 2016). Exploring these and other accounts of causal judgment in overdetermination and preemption remains an important area for further research.

iii. Future directions

Our model predicts the form that actual causal judgments will take in a variety of causal structures. Here, we tested its predictions in a two-variable conjunctive structure; future work should investigate the quantitative form that people's causal judgments take in other structures (e.g. structures with more variables, or that combine conjunctive and disjunctive relationships).

Perhaps the most important prediction of our model is that people will use their judgments of actual causation to help choose interventions. This could be tested by presenting people with a novel causal system where they can observe past episodes, make causal judgments, and then intervene to produce desired outcomes. Our model predicts that a person's average judgment of the extent to which a variable C caused an outcome E will predict how often they choose C as an intervention to produce E.

Additionally, as noted above, our model admits of at least two interpretations. It could be interpreted as describing only the function of actual causal judgments—Marr's computational level of analysis (Marr, 1982). On this view, the computation being performed by actual causation results in an approximation of intervention effectiveness, but the algorithm used to perform that computation may look

[24] This explanation assumes that there are restrictions on which circumstances are permissible to modify. For details, see Halpern and Hitchcock (2011).

[25] This account would offer a functional explanation for a curious empirical result. When generating the additional observation, people would presumably be influenced by the base rates of the antecedent events. Concretely, if the first soldier is more likely to shoot than the second, people should be more likely to generate an additional sample where the second soldier didn't shoot—and hence, surprisingly, should consider the *more common* factor more causal. This is exactly what occurs in overdetermined cases (Icard et al., 2017).

different than what we have proposed. For instance, the basic functional design embodied by our model may be consistent with quite a different algorithm for generating sample-based estimates of intervention effectiveness, such as an algorithm that samples counterfactual scenarios (Gerstenberg, Goodman, et al., 2015; Hitchcock & Knobe, 2009; Icard et al., 2017; Kominsky et al., 2015; Phillips et al., 2015). Alternatively, our model could be interpreted as a mechanistic description of how people make actual causal judgments: They first judge whether a variable was necessary, then apply the correction term, and so on (Marr's algorithmic level of analysis). Refining our understanding of the particular algorithm that people employ during causal judgment is an important goal for future research.

VII. Conclusion

Actual causation is a central concept in people's everyday lives, but its form and function have remained mysterious. Here, we find that, if actual causation is designed to approximate the effectiveness of potential interventions, it will take a form that matches observed empirical patterns. The nature of causal concepts may therefore be illuminated by focusing on their function (Woodward, 2014).

Acknowledgements

We thank Jonathan Kominsky for sharing his data on supersession, and the Moral Psychology Research Lab for their advice and assistance. AM and FC were supported by Grant N00014-14-1-0800 from the Office of Naval Research to FC. TG was supported by the Center for Brains, Minds & Machines (CBMM), which is funded by the National Science Foundation's Science and Technology Center (Award CCF-1231216).

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological bulletin*, *126*(4), 556.

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *cognition*, *167*, 25–37.

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2017, aug). Stability, breadth and guidance. *Philosophical Studies*. Retrieved from `https://doi.org/10.1007%2Fs11098-017-0958-6` doi: 10.1007/s11098-017-0958-6

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

Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (accepted). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, *58*(4), 545–567.

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.

Chockler, H., Halpern, J. Y., & Kupferman, O. (2008). What causes a system to satisfy a specification? *ACM Transactions on Computational Logic*, *9*(3), 20.

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology*, *79*, 102–133.

Collingwood, R. G. (2014). *An Essay on Metaphysics*. Martino Fine Books.

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

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuro-*

science, *8*(12), 1704.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

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

Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *bioRxiv*, 196923.

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2012). Noisy newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Gerstenberg, T., Goodman, N., Lagnado, D., & Tenenbaum, J. (2014). From counterfactual simulation to causal judgment. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 788–793). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & 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, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from `https://doi.org/10.1177%2F0956797617713053` doi: 10.1177/0956797617713053

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

Glynn, L. (2017). A proposed probabilistic extension of the Halpern and Pearl definition of actual causation. *British Journal for the Philosophy of Science*, *68*, 1061–1124.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

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.

Halpern, J. Y. (2013). From causal models to counterfactual structures. *The Review of Symbolic Logic*, *6*(2), 305–322.

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

Halpern, J. Y., & Hitchcock, C. (2011, June). Actual causation and the art of modeling. *arXiv:1106.2652 [cs]*. Retrieved from `http://arxiv.org/abs/1106.2652` (arXiv: 1106.2652)

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.

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

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. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*(6), 273–299.

Hitchcock, C. (2007). Three concepts of causation. *Philosophy Compass*, *2*(3), 508–516.

Hitchcock, C. (2009). Structural equations and causation: six counterexamples. *Philosophical Studies*, *144*(3), 391–401.

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

Hitchcock, C. (2017). Actual causation: What's the use? *Making a Difference: Essays on the Philosophy of Causation*, 5.

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

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80–93. Retrieved from https://doi.org/10.1016%2Fj.cognition.2017.01.010 doi: 10.1016/j.cognition.2017.01.010

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1–17.

Kahneman, D., & 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.

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

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.

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

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197. Retrieved from http://www.jstor.org/stable/2678389

Lombrozo, T. (2006, October). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364661306002117 doi: 10.1016/j.tics.2006.08.004

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

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

Luhmann, C. C., & Ahn, W.-k. (2005, July). The Meaning and Computation of Causal Power: Comment on and. *Psychological review*, *112*(3), 685–707. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2677809/ doi: 10.1037/0033-295X.112.3.685

Mackie, J. L. (1974). *The cement of the universe.* Oxford: Clarendon Press.

Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of personality and social psychology*, *71*(3), 450.

Markant, D. B., & Gureckis, T. M. (2014). Is it better to select or to receive? learning via active and passive hypothesis testing. *Journal of Experimental Psychology: General*, *143*(1), 94.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc.

McGill, A. L., & Tenbrunsel, A. E. (2000). Mutability and propensity in causal selection. *Journal of Personality and Social Psychology*, *79*(5), 677-689. Retrieved from http://dx.doi.org/10.1037/0022-3514.79.5.677 doi: 10.1037/0022-3514.79.5.677

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

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.

Pearl, J. (1999). Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, *121*(1-2), 93–149.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge University Press.

Perales, J. C., & Shanks, D. R. (2007, aug). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*(4), 577–596. Retrieved from http://dx.doi.org/10.3758/bf03196807 doi: 10.3758/bf03196807

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*(18), 4649–4654.

Phillips, J., & Kominsky, J. (2017). Causation and norms of proper functioning: Counterfactuals are (still) relevant. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 931–936). Austin, TX: Cognitive Science Society.

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

Rottman, B. M., & Hastie, R. (2013). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin*. Retrieved from http://dx.doi.org/10.1037/a0031903 doi: 10.1037/a0031903

Samland, J., & Waldmann, M. R. (2016). How prescriptive norms influence causal inferences. *Cognition*, *156*, 164–176.

Sartorio, C. (2012, Dec). Two wrongs do not make a right: Responsibility and overdetermination. *Legal Theory*, *18*(04), 473-490. Retrieved from http://dx.doi.org/10.1017/S1352325212000122 doi: 10.1017/S1352325212000122

Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science*, *67*(2), 285–300.

Schaffer, J. (2004). Causes need not be physically connected to their effects: The case for negative causation. In C. R. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 197–216). Blackwell.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*(2), 241–263.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *nature*, *529*(7587), 484–489.

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

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

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247. Retrieved from http://dx.doi.org/10.1146/annurev-psych-010814-015135 doi: 10.1146/annurev-psych-010814-015135

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

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323–348.

Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual (" but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems*,

64(4), 241–264.

Stephan, S., & Waldmann, M. R. (2016). Answering causal queries about singular cases. In *Proceedings of the 38th annual conference of the cognitive science society* (pp. 2795–2801).

Stephan, S., Willemsen, P., & Gerstenberg, T. (2017). Marbles in inaction: Counterfactual simulation and causation by omission. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1132–1137). Austin, TX: Cognitive Science Society.

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

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). MIT press Cambridge.

Tian, J., & Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, *28*(1-4), 287–313.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2016). Stable causal relationships are better causal relationships. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2663–2668). Austin, TX: Cognitive Science Society.

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

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *19*(3), 231.

Wells, G. L., & Gavanski, I. (1989). Mental simulation of causality. *Journal of personality and social psychology*, *56*(2), 161.

Willemsen, P. (2016). Omissions and expectations: A new approach to the things we failed to do. *Synthese*, 1–28.

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

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

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

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

Woodward, J. (2007). Interventionist Theories of Causation in Psychological Perspective.. Retrieved 2017-11-03, from `http://philsci-archive.pitt.edu/3132/`

Woodward, J. (2014, December). A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment). *Philosophy of Science*, *81*(5), 691–713. Retrieved from `https://www-journals-uchicago-edu.ezp-prod1.hul.harvard.edu/doi/abs/10.1086/678313` doi: 10.1086/678313

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

VIII. Appendix

All data and analysis code can be found at `https://osf.io/j8swe/`.

i. Why measures of intervention effectiveness should condition on E = 0

Our definition of the general effectiveness of an intervention — $GE(C \rightarrow E)$ — conditions on the absence of the outcome; it is defined as $Prob(E_{C=1} = 1 \mid E = 0)$. In other words, the usefulness of an intervention to bring about an outcome is measured by situations where the outcome is currently absent. The purpose of this section is to elaborate on why we defined intervention effectiveness this way. To reiterate, the argument is that, in decision-making situations where a person is trying to bring about a goal ($E = 1$), the one thing that the person almost always knows is that the goal does not currently obtain (i.e. before the intervention, $E = 0$). If she knows that the goal already obtains, then she would not be trying to bring it about, and so would not consult her representation of the general effectiveness of potential interventions. Since the representation is only used in situations where $E = 0$, it is beneficial to condition the representation on that knowledge. Otherwise, distortions can occur. We explore these distortions in two examples.

In one type of distortion, a perfectly effective intervention can be represented as ineffective. Imagine a lightbulb that turns on if and only if two switches are in the same orientation (up or down; Figure 8A). Intuitively, if you want to turn the light on, each switch is a perfectly effective intervention. This is because, in situations where you want to turn the light on, the light is always currently off – and so flipping either switch is guaranteed to turn the light on.

This intuition is captured by the general intervention effectiveness of the switch (say, `S_1`) on the lightbulb (`LB`), $Prob(\mathtt{LB}_{\mathtt{S\_1}=1} = 1 \mid LB = 0)$. Conditioned on the light being off ($LB = 0$), there are only two possible joint settings of the exogenous variables: the first switch is up and the second down, or the reverse. Either way, flipping either switch is guaranteed to turn the light on. Hence, the general effectiveness is one; $Prob(LB_{S_1=1} = 1 \mid LB = 0) = 1$.

But if you did not condition on the light currently being off, then you would consider each switch to be a less effective intervention, because you would be mistakenly incorporating cases where the lightbulb is currently on (and hence the switch turns the light off). Imagine that our model of general effectiveness was $Prob(LB_{S_1=1} = 1)$ – the same value, without conditioning on the absence of the outcome. Then, we would consider the effectiveness of the switch to be substantially less than one. This is the distortion caused by considering cases, like when the light is already on, that are not relevant for the person's decision-making.

Another type of distortion occurs when an utterly ineffective intervention is represented as effective. Imagine the following causal system. A prisoner is being put to death by firing squad. The firing squad has two soldiers, who will both shoot if and only if their commander orders them to. Both bullets are necessary to kill the prisoner; if only one soldier fires, the prisoner will live.

Suppose you want the prisoner to die. How effective is it to intervene on one of the soldiers? Intuitively, it is not effective at all. The only case when your actions are worth taking is when the prisoners still alive. But in this case, the commander has not ordered them to shoot ($CO = 0$)—so intervening on one of the soldiers will do nothing.

This intuition is captured by the general intervention effectiveness of the first soldier (S_1) on the prisoner's death (PD), $Prob(PD_{S_1=1} = 1 \mid PD = 0)$. Given $PD = 0$, the only possible value of the exogenous variable CO is 0. Hence, in the mutilated model where $S_1 = 1$, the probability that $PD = 1$ is zero. In other words, intervening on one soldier has zero effectiveness.

In contrast, if we don't condition on $PD = 0$, then we end up mistakenly assigning effectiveness to intervening on the soldier. If our model of general effectiveness was $Prob(PD_{S_1=1})$, the effectiveness of the soldier would be the proba-

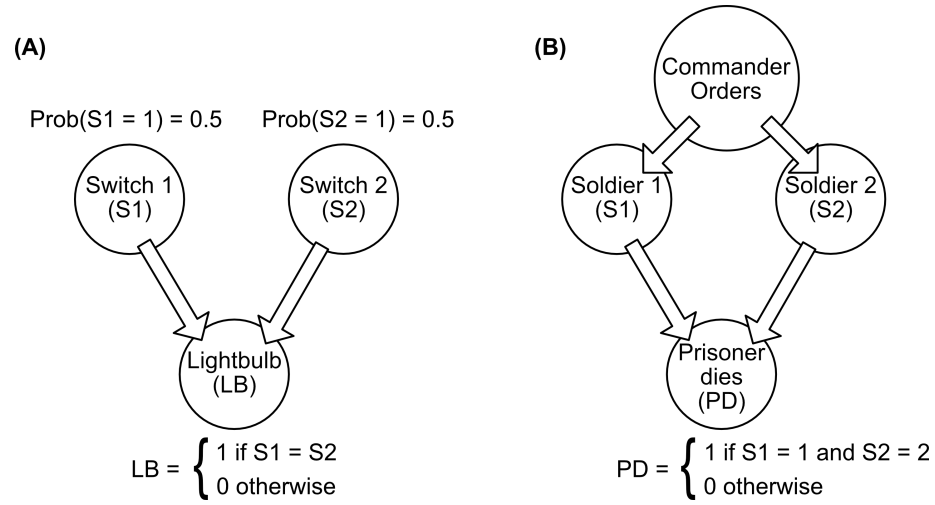


Figure 8: *(A) A causal system in which, intuitively, an intervention is perfectly effective. (B) A causal system in which, intuitively, an intervention is perfectly ineffective. In both examples, we capture these intuitions by conditioning on the absence of the outcome.*

bility that the prisoner is dead in the mutilated model where S_1 is fixed to 1 – which is just $Prob(CO = 1)$, the prior probability that the commanding officer gives the order to shoot. Clearly, this is an inadequate notion of effectiveness in this case.

These examples illustrate the motivation to condition on $E = 0$ when calculating the general effectiveness of potential interventions. Note that, although these cases have extreme values for intervention effectiveness (i.e. 0 or 1 for all interventions), the causal ratings predicted by the SAMPLE model are still nontrivial. Future research should test whether people's causal judgments in these cases conform to the SAMPLE model.

One final point: You might wonder, why don't we condition on the absence of the potential intervention also? In other words, why not condition on $C = 0$? The reason is that, across all decision situations, it is not generally true that $C = 0$. For instance, it is almost always the case that a student trying to get into college has not already been accepted; we can assume $E = 0$. But we cannot assume that the student has not already improved their writing, or joined the debate team, or pulled any of the scholastic levers that could serve as potential interventions. Decision situations will vary wildly in which potential interventions have already been employed, and hence the representation of the general, situation-agnostic effectiveness of an intervention should not condition on the intervention being absent.

ii. Proof of Proposition 1

In this section, we prove the relationship stated in Proposition 1: that the general effectiveness of intervening on C to produce E can be estimated by determining necessity in cases where E is present, and applying a correction term. Here, instead of using $GE(C \rightarrow E)$ to denote the general effectiveness, we will use its more common name: the probability of enablement, denoted $PE(C, E)$. To restate it formally:

$$PE(C, E) \approx \frac{1}{k} \sum_{i=1}^{k} AC(C, E, u_i)$$

$$= \frac{1}{k} \sum_{i=1}^{k} I(f_E^{C=0}(u_i) = 0) * \frac{Prob(C = 0)}{Prob(C = 1)} \frac{Prob(E = 1)}{Prob(E = 0)}$$

The proof draws heavily on Pearl (1999). Consider a causal system with at least two variables, C and E. We have already defined the probability of enablement for C on E as

$PE(C, E) = Prob(E_{C=1} = 1 \mid E = 0)$. It will be useful to define one other value: the probability of disablement for C on E, or $PD(C, E)$. This is, intuitively, the opposite of the probability of enablement. It is the probability that, given that the outcome E is present, turning C off would turn E off. Formally:

$$PD(C, E) = Prob(E_{C=0} = 0 \mid E = 1) \qquad (1)$$

Intuitively, the probability of disablement that can be estimated by assessing the necessity of C in situations where $E = 1$. We state this formally.

Proposition 2. Suppose $u_1, ..., u_k$ are randomly sampled observations from the joint probability distribution over all exogenous variables, given $E = 1$. As k increases, the following holds:

$$PD(C, E) \approx \frac{1}{k}\sum_{i=1}^{k} I(f_E^{C=0}(u_i) = 0)$$
$$= \frac{1}{k}\sum_{i=1}^{k} I(C \text{ was necessary for } E \text{ in } u_i)$$

(The proof follows immediately from the law of large numbers.)

This is, of course, precisely the sampling procedure that we want to use. Our task, then, is to relate the probability of enablement (our target) to the probability of disablement (which can be estimated via this procedure). This is accomplished by one crucial assumption. PE and PD can be easily transformed into each other, so long as the effect on E from intervening on C is not dependent on the current value of C. In other words, $Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = c)$ must equal $Prob(E_{C=1} = 1, E_{C=0} = 0)$. This is the "no-confounding", or "exogeneity" assumption (Pearl, 2000), which we discussed above.

If the no-confounding assumption holds, then the probabilities of enablement and disablement are related in the following way (Pearl, 1999).

Proposition 3. Suppose that $Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = c) = Prob(E_{C=1} = 1, E_{C=0} = 0)$. Then:

$$PE(C, E) = \frac{P(C = 0)P(E = 1)}{P(C = 1)P(E = 0)} PD(C, E)$$

Proof. By the law of total probability,

$$PE(C, E) = Prob(E_{C=1} = 1 \mid E = 0)$$
$$= Prob(E_{C=1} = 1 \mid E = 0, C = 0) * Prob(C = 0 \mid E = 0)$$
$$+ Prob(E_{C=1} = 1 \mid E = 0, C = 1) * Prob(C = 1 \mid E = 0)$$

We will make use of the fact that $Prob(E_{C=c} = e \mid E \neq e, C = c) = 0$. In words, if C is already equal to a value c, then intervening to set it to c cannot change the value of E (Pearl, 2000). (Another way to write this is, used below, is that $Prob(E = e, C = c) = Prob(E_{C=c} = e, C = c)$.) Hence:

$$PE(C, E) =$$
$$Prob(E_{C=1} = 1 \mid E = 0, C = 0) * Prob(C = 0 \mid E = 0) =$$
$$Prob(E_{C=1} = 1, E = 0, C = 0) * \frac{Prob(C = 0 \mid E = 0)}{Prob(E = 0, C = 0)} =$$
$$Prob(E_{C=1} = 1, E = 0, C = 0) * \frac{1}{Prob(E = 0)} =$$
$$Prob(E_{C=1} = 1, E_{C=0} = 0, C = 0) * \frac{1}{Prob(E = 0)} =$$
$$Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = 0) * \frac{Prob(C = 0)}{Prob(E = 0)}$$

By identical logic:

$$PD(C, E) =$$
$$Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = 1) * \frac{Prob(C = 1)}{Prob(E = 1)}$$

By the no-confounding assumption,

$$Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = 0) =$$
$$Prob(E_{C=1} = 1, E_{C=0} = 0 \mid C = 1)$$

Hence,

$$PE(C, E) * \frac{Prob(E = 0)}{Prob(C = 0)} = PD(C, E) * \frac{Prob(E = 1)}{Prob(C = 1)}$$

and

$$PE(C, E) = \frac{P(C = 0)P(E = 1)}{P(C = 1)P(E = 0)} PD(C, E)$$

□

Combining Propositions 2 and 3 completes the proof of Proposition 1.

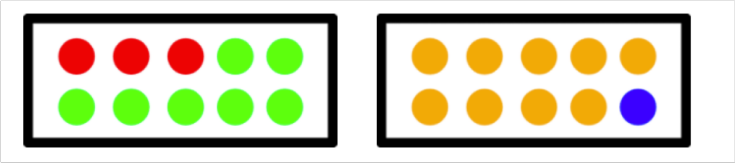
iii. Full vignette

See Figure 9.

iv. Three-dimensional visualizations of model predictions and results

See Figure 10.

Joe is playing a casino game. There are two boxes in front of him. Each box has some colored balls inside it, which are going to get shuffled together.



Joe is going to close his eyes, reach in, and randomly choose a random ball from each box. He will win a dollar if he chooses a **green ball** from the first box **AND** a **blue ball** from the second box.

Joe closes his eyes, reaches in, and chooses a green ball from the first box and a blue ball from the second box. So Joe wins a dollar.

Please tell us much you agree or disagree with this statement: Joe's first choice (where he chose a green ball from the first box) caused him to win the dollar.

Figure 9: *Full text of the vignette in the behavioral experiment. In the example,* $Prob(\texttt{green} = 1) = .7$ *and* $Prob(\texttt{blue} = 1) = .1$.

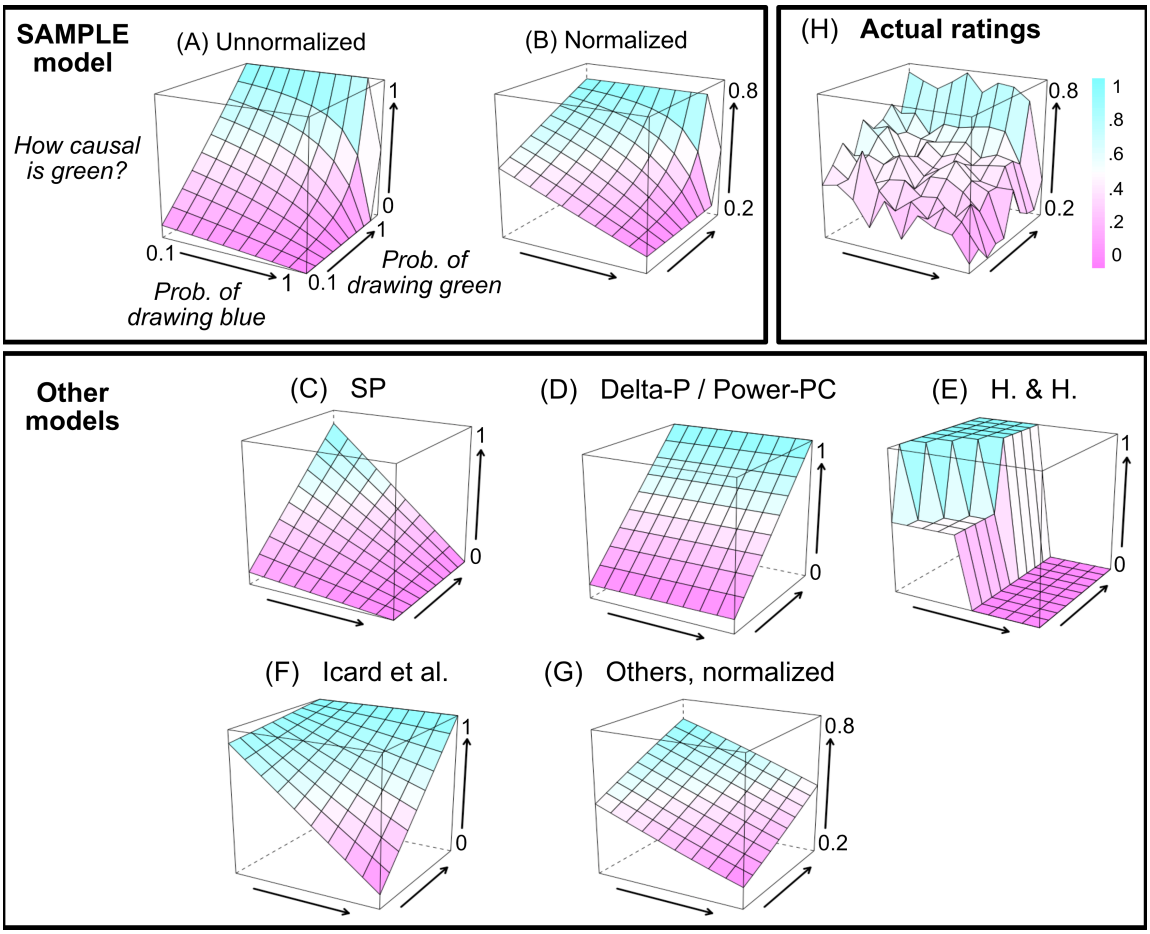


Figure 10: *A three-dimensional visualization of the model predictions and results in our experiment.*