Counterfactuals and the logic of causal selection

Tadeg Quillien and Christopher G Lucas

School of Informatics

University of Edinburgh

Author Note

This is the last submitted version of a paper published in *Psychological Review*.

Open Science Practices. Data and analysis code for all studies are available at https://osf.io/h42f7/. Studies 2 to 4 were pre-registered (Study 2a: https://osf.io/6cg4r/; Study 2b: https://osf.io/qv5g4; Study 3: https://osf.io/dfz5p; Study 4: https://osf.io/q3xa4/).

Prior dissemination A talk based on this research was presented at the 48th meeting of the Society for Philosophy and Psychology. This preprint is available at https://psyarxiv.com/ts76y/.

We have no conflicts of interest to report. Correspondence concerning this article should be addressed to Tadeg Quillien, School of Informatics, University of Edinburgh. E-mail: tadeg.quillien@gmail.com

Abstract

Everything that happens has a multitude of causes, but people make causal judgments effortlessly. How do people select one particular cause (e.g. the lightning bolt that set the forest ablaze) out of the set of factors that contributed to the event (the oxygen in the air, the dry weather...)? Cognitive scientists have suggested that people make causal judgments about an event by simulating alternative ways things could have happened. We argue that this counterfactual theory explains many features of human causal intuitions, given two simple assumptions. First, people tend to imagine counterfactual possibilities that are both a priori likely and similar to what actually happened. Second, people judge that a factor C caused effect E if C and E are highly correlated across these counterfactual possibilities. In a reanalysis of existing empirical data, and a set of new experiments, we find that this theory uniquely accounts for people's causal intuitions.

Keywords: Causal selection, Causation, Counterfactuals, Computational modeling

Counterfactuals and the logic of causal selection

Most things that happen have a dizzying number of causes. Consider two cars crashing at an intersection. The accident was contingent on the fact that both drivers crossed the intersection, that neither driver had been delayed in traffic earlier, that someone invented the automobile, and so on. Of course, people spontaneously say that the driver crossing at the red light caused the accident — just like they say that the victory in the swing-state caused the presidential candidate's success, or that the last-second shot caused the basketball team's victory (Alicke et al., 2011; Knobe & Fraser, 2008; Quillien & Barlev, 2022; Henne, Kulesza, et al., 2021).

Causal selection — the ability to mentally foreground one of the many causes of an event — plays a key role in our cognitive lives. It enables us to give effective explanations (Kirfel et al., 2021), assign blame and praise (Alicke et al., 2011), plan for the future (Morris et al., 2018), and identify factors that may lead to the same outcome in similar future situations (Woodward, 2006; Lombrozo, 2010; Hitchcock, 2012). Our intuitive concept of causation is the foundation for many other concepts, such as intentional action (Quillien & German, 2021) and responsibility (Gerstenberg et al., 2018). How we select causes has preoccupied philosophers since at least the time of John Stuart Mill (1843), but it also has consequential real-world implications, in settings ranging from law (Hart & Honoré, 1985; Knobe & Shapiro, 2021) to industrial safety (Hanley, 2021). But despite the ubiquity and subjective simplicity of causal selection, its underlying cognitive processes are not yet fully understood by cognitive scientists.

This paper offers a theory of how people judge the relative importance of the causes that led to an event. In broad strokes, our theory holds that when people judge whether event C caused outcome E, they do the following:

a) They consider alternative possibilities for how the situation could have unfolded. They tend to imagine alternative possibilities that are a priori likely, but that also do not deviate too much from what actually happened.

b) They compute an 'effect size' measure that quantifies how much C tends to influence E across all these imagined possibilities. In many cases, this consists in computing the correlation between C and E across all these alternative possibilities. People think that C caused E if this measure of effect size is large.

Our account holds that causal judgment relies on counterfactual reasoning. This idea has already been shown to explain many aspects of people's causal intuitions (Icard et al., 2017; Quillien, 2020; Gerstenberg et al., 2021). Yet there has been surprisingly little connection between theories of causal judgment and formal models of counterfactual reasoning.

We bridge this gap. Our theory combines an empirically-grounded formal model of counterfactual reasoning (Lucas & Kemp, 2015) with a model of how people compute how much a given event counts as the cause of an effect (Quillien, 2020). Together, these two relatively simple building blocks parsimoniously account for people's causal intuitions.

In this introduction, we first present the problem of causal selection[1], and give an intuition for why our theory works the way it does. Then we lay out the building blocks of

[1] We use the term 'causal selection' in an abstract sense, to refer to the operation by which the mind judges some factors to be more causally responsible than others for an outcome, when it is known that all these factors had a causal influence on the outcome. That is, we talk about causal *selection* to distinguish our problem from related problems such as causal inference. In particular we do not mean to use the term to refer to a particular dependent variable or behavior (where people make a discrete choice of one cause among an array of candidates). Indeed we collect judgments by asking participants to make graded ratings for each of the causes in a particular scenario (instead of a discrete choice). This is justified on theoretical grounds: following existing work (e.g. Morris et al., 2018; Kominsky et al., 2015; Lagnado et al., 2013; Morris et al., 2019; O'Neill, Henne, Bello, et al., 2022; Danks, 2017), we assume that people make causal judgments by computing a graded score of the 'actual causal strength' of a given candidate cause. Note that this notion of 'actual' causal strength differs from the notion of causal strength used in theories of how people reason about general causal relationships (e.g. Cheng, 1997). For philosophical debates about whether causation really comes in degrees see Kaiserman (2016, 2018), Sartorio (2020), and Demirtas (2022).

the theory. On the basis of past work on counterfactual reasoning, we identify one plausible model for how people might sample alternative possibilities. Then we describe a model for how the mind might use these alternative possibilities to make causal judgments.

We then test our account against existing empirical data. We find that data from past studies are consistent with our proposal, but that these studies do not discriminate well between our account and some alternatives. We then report data from four new experiments that suggest that our proposal is a better fit to people's causal judgments than these alternatives.

The problem of causal selection

Scope of the problem

A large literature on causal cognition has explored how people learn causal facts about the world (e.g. Cheng, 1997; Gopnik et al., 2004; Griffiths & Tenenbaum, 2005, 2009; Lucas & Griffiths, 2010; Bramley et al., 2015; Bramley et al., 2017; Zhao et al., 2021) or how they infer whether an event would have happened in the absence of another event (Gerstenberg et al., 2021; Kelley, 1973; Ahn et al., 1995; Stephan et al., 2020). Here we are concerned with a different problem. Even when we know everything there is to know about a causal system, it is still not obvious (from an information-processing perspective) how we should describe why something happened (Mill, 1843; Hart & Honoré, 1985; Hesslow, 1988).

For example, suppose we know that fires occur when a spark ignites a flammable material and there is oxygen in the air to fuel the combustion. We also know that today a lightning bolt struck a dead tree, there was oxygen in the air, and the forest caught fire. Was it the lightning bolt or the oxygen that caused the fire?

We have the intuition that the lightning bolt was *the* cause of the fire, but this intuition does not follow transparently from our knowledge of the relevant facts. One can in principle imagine a rational agent who learns a complete causal model of the situation, yet does not have any intuition about whether one factor was 'more of a cause' than the

other. Here we are interested in the computations that generate these kinds of intuitions. Therefore we will restrict our attention to situations where there is no learning problem to be solved, and people already know the relevant facts about the causal system.

Our problem is also different from the problem of making *categorical* causal judgments. Although people think that the lightning bolt is 'more of a cause' than the oxygen, there is still a sense in which the oxygen clearly had a causal influence on the fire. After all, without oxygen in the air the fire could not have started. By contrast, many other factors (such as the color of the trees in the forest) were clearly non-causal. So, people make a categorical distinction between those factors that had a causal influence on the fire (the oxygen in the air, the lightning bolt that started the blaze...), and those that did not (the color of the trees, the fact that it was Tuesday...). Some theories of causation are designed to formalize how we make (or should make) these categorical distinctions (Halpern & Pearl, 2005; Halpern, 2016; Hitchcock, 2001, see also the 'first step' in Gerstenberg et al., 2021)[2].

It is useful to view causal selection as what happens after this process of categorical causal judgment has played out[3]. After we have identified some factors (the lightning bolt, the oxygen, the dry leaves on the ground...) as causes of the fire, the process of causal selection determines which factors we single out as the important ones.

---

[2] Some classic problems about causation are best analysed as concerning categorical causal judgment rather than causal selection. For example, in a case of 'causal pre-emption', the pre-empted variable clearly is thought to have had zero causal influence on the outcome, and so it is (probably) not even flagged as a candidate for the kinds of computation we describe in this paper. Similarly, the principle that an event that does not occur cannot be a cause is a basic assumption of models of categorical causal judgment.

[3] Doing so is conceptually useful from a task-analysis perspective – we are not making a claim about whether the processes are really implemented sequentially.

The function of causal selection

Our theory belongs to a general framework that views causation as a matter of counterfactual dependence (Woodward, 2006; Halpern & Pearl, 2005; Lewis, 1973a; Woodward, 2003; Gerstenberg et al., 2017; Gerstenberg, 2022; Wells & Gavanski, 1989; Krasich et al., n.d.). According to counterfactual theories, "C caused E" means (roughly) that if C had not happened, then E would not have happened either. Presented in this way, counterfactual theories seem to suggest that people only consider counterfactuals where the value of C is changed. But our theory assumes that people actually simulate several alternative possibilities, where not only the value of C, but that of other variables, can change.

Why might it be sensible to simulate several alternative possibilities rather than just one? We argue that this design feature makes sense when we consider the function of causal selection.

Many philosophers and cognitive scientists think that cognitive mechanisms for causal selection are designed for identifying causes that are *generalizable* (sometimes also called *robust, exportable, insensitive*, or *invariant*; Morris et al., 2018; Woodward, 2006; Lombrozo, 2010; Woodward, 2021; Hitchcock and Knobe, 2009). Generalizable causes are causes that led to an effect in a way that did not depend on overly idiosyncratic details of the situation. For instance, oxygen is not a generalizable cause of the fire, because there are many situations in which the presence of oxygen would not lead to a fire. By highlighting generalizable causes, causal selection helps us make effective interventions and reliable predictions in future situations (Morris et al., 2018; Hitchcock, 2012; Hanley, 2021).

To assess whether oxygen was a robust cause of the forest fire, it is not enough to simply imagine a counterfactual situation where everything is the same as in the actual situation except there is no oxygen in the air. We must additionally imagine situations where other features of the situation are different — for example situations in which there was no lightning bolt. Only when we imagine such situations does it become apparent that

oxygen is not a robust cause of the fire.

In sum, a general hypothesis about the function of causal selection suggests that, when people make causal judgments, they simulate several different alternative possibilities. We will assume that these alternative possibilities are an input to a computation of the "causal strength" of the candidate cause.

In what follows we flesh out this proposal, by describing formal models of how people might sample alternative possibilities and compute causal strength.

Causal models

Our approach is grounded on the idea that much of human cognition consists in operations over probabilistic causal models of the world (Chater & Oaksford, 2013; Sloman & Lagnado, 2015; Gerstenberg & Tenenbaum, 2017; Lake et al., 2017; Pearl & Mackenzie, 2018). A probabilistic causal model represents a given aspect of the world in the form of variables, causal relationships between them, and probability distributions (Pearl, 2000).

As a simple example, consider a game where you randomly draw a ball from each of two boxes (Figure 1a). The first box contains 1 green ball and 9 black balls, and the second box contains 4 blue balls and 6 black balls. The rules of the game are that you win if and only if you draw a colored ball from both boxes. We can represent how the game works using the probabilistic causal model shown in Figure 1B:

The causal model represents the world in terms of variables[4]. For instance, the variable *Green* takes value 1 if the ball drawn from the first box is green and value 0 otherwise. The *Win* variable is called an endogenous variable, because its value depends on the value of the two variables from which it receives arrows. By contrast, the *Green* and

[4] More technically, here we use the formalism of Functional Causal Models (sometimes called Structural Equation Models) to model how people represent the causal structure of the world (Pearl, 2000). There are other formalisms for causal modeling, for example Causal Bayes Nets. We use Functional Causal Models because, as argued by Pearl, they are more expressive when it comes to representing counterfactual possibilities.

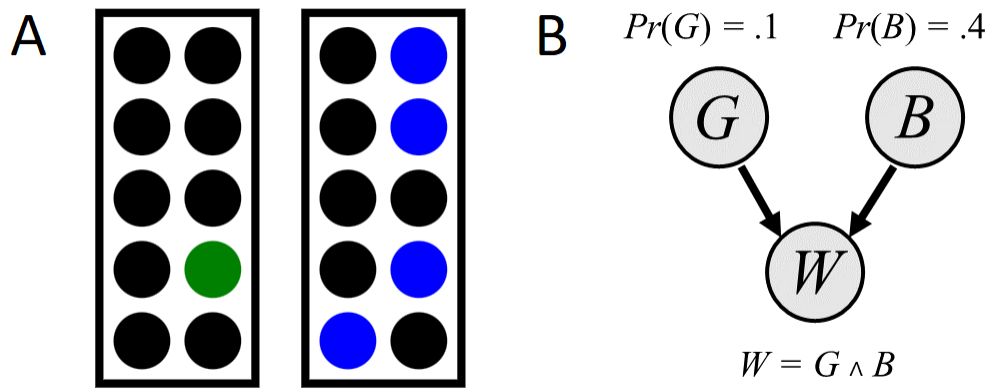

Figure 1

A) *The two boxes from which a player draws randomly, in the game of chance we will use as our running example (inspired by Morris et al., 2019).* ***B)*** *A probabilistic causal model of the game. The variables G and B (for Green and Blue) each have an associated probability distribution, corresponding to the proportion of colored balls in the relevant box. The variable W (for Win) is determined by the structural equation $W = G \wedge B$: it takes value 1 if and only if the player draws both a green ball and a blue ball.*

Blue variables are called exogenous variables: their value does not depend on the value of other variables. Instead, their value has a probability distribution: for example, because there is 1 green ball out of the 10 balls in the first box, the variable *Green* has prior probability .1. See the Supplementary Information (https://osf.io/vnh84) for a brief explanation of causal models as they relate to the current work, and Halpern, 2016 or Pearl, 2000 for a detailed treatment.

Our proposal is that when people make causal judgments – for instance when they judge whether drawing a green ball caused a player to win the game – they simulate counterfactual possibilities by *sampling* from their causal model of the situation.

Sampling from causal models

Many useful operations on causal models consist in sampling possibilities from them (Goodman et al., 2014; Icard, 2016; Sanborn & Chater, 2016)[5]. For instance, to compute the average probability that a player would win the game described above, we can simulate many possible rounds of the game and compute the proportion of simulations where the player wins the game.

To simulate a possibility from a causal model, we first sample a value for each exogenous variable, and then we set the value of the other variables in the model as a function of the value of the exogenous variables, and the causal relationships encoded in the model. For example, to simulate one possible way a round of the game could unfold, we could sample a value for *Green* according to its prior probability (i.e., with probability .1 we imagine that the player draws a green ball, with probability .9 we imagine that he draws a black ball) and do the same for the value of *Blue*. Then we would set the value of *Win* as a function of the rules of the game and the values of *Green* and *Blue* (if we imagined the player drawing a green ball and a blue ball, we imagine that he won the game).

[5] In computational cognitive science, the concept of sampling is invoked for at least two different purposes. First, exact inference is often not computationally tractable, meaning that many inference problems must be solved by approximation methods, and many of these methods (e.g. Markov Chain Monte Carlo) rely on sampling (MacKay, 2003). There is evidence that the human mind might use such sample-based approximation methods (Sanborn & Chater, 2016; Vul et al., 2014; Lieder et al., 2018; Zhu et al., 2020). Second, regardless of whether the human mind actually uses sampling-based approximation, it is often conceptually useful to describe computations in terms of sampling. For instance, one way to define a probability distribution is to write a probabilistic program, and imagine that we take an infinite number of samples from that program (Goodman et al., 2014; Ackerman et al., 2011). Here we are mostly inspired by the latter use of the concept of sampling. We think that the easiest way to formulate our theory is to describe a sampling process, but we are not committed to the claim that people actually draw samples when they make causal judgments. As an analogy, it is possible to predict how someone will answer a question about probability by assuming that they draw samples, even if the person actually consults the equations of probability theory.

Note that for solving some problems – like computing the probability that a player wins the game – it seems very reasonable to sample the value of a variable in exact proportion to its real-world probability when we simulate possibilities from a causal model. There is a 10% probability that a player draws a green ball from the first box, so a natural way to obtain correct results in our calculation is to sample $Green = 1$ with probability .1. But for some problems, people might also sample the value of a variable with a different probability than its real-world probability. For some purposes we may want to sample the value $Green = 1$ with probability .5, even though the objective probability of drawing a green ball is .1. We call *sampling propensity* the probability with which one samples a variable from a causal model (Icard, 2016). By contrast, when we talk about a variable's probability, we refer to our best available estimate of the probability of the event that the variable represents.

Here we are especially interested in how people simulate *counterfactual* possibilities. To simulate counterfactual possibilities, people first consider something that happened and then imagine other possible ways that this particular situation could have unfolded. How do people simulate such counterfactuals when they make causal judgments? One promising way to look for an answer is to look at how people reason about counterfactuals in other domains.

A formal model of counterfactual sampling

To make things concrete, suppose our friend Alice played one round of the game, drew a green ball from the first box, a blue ball from the second box, and (since she got two colored balls) won the game. How do people simulate alternative ways that the game could have happened?

A very simple hypothesis would be that people simulate counterfactual possibilities according to their prior probability. That is, people might simulate counterfactual possibilities by sampling exogenous variables from their probability distributions. Under

that hypothesis, when people imagine other possible outcomes of the round of the game played by Alice, they are just doing the same thing they do when they imagine what the average possible round of the game would look like.

Intuitively, the hypothesis fails to capture something fundamental about counterfactual thinking. Counterfactual thinking is not just concerned with the general probability of an event. Instead, it is about the different ways that one particular situation could have unfolded. Counterfactuals tend to be 'close' to (or 'centered on') the actual world (Lewis, 1973b; Stalnaker, 1981; Hiddleston, 2005; Pearl, 2013; De Brigard et al., 2021; Stanley et al., 2017)[6]. But on the very simple hypothesis, people totally disregard what actually happened when they simulate counterfactuals. Consider that, because there are fewer than 50% of colored balls in each urn, the most likely a priori outcome of the game is one where Alice draws two black balls. But, given that she actually drew two colored balls, it seems somewhat odd to think of her drawing two black balls as the most natural alternative to what actually happened.

Research on counterfactual thinking actually suggests that people simulate counterfactuals that are both *likely* and *close to what actually happened* (Lucas & Kemp, 2015, 2012). The relevant evidence comes from studies on how people reason about counterfactual conditionals (e.g. Lucas & Kemp, 2015; Sloman & Lagnado, 2005; Over et al., 2007; Rips, 2010; Rips & Edwards, 2013; Dehghani et al., 2012; Oaksford & Chater, 2007; Gerstenberg et al., 2013; Pfeifer & Tulkki, 2017; Byrne & Johnson-Laird, 2020; Skovgaard-Olsen et al., 2021). Counterfactual conditionals are statements such as "if Alice had drawn a black ball from the first box, she would have lost the game" (Lewis, 1973b; Stalnaker, 1981; Hiddleston, 2005; Lassiter, 2017; Starr, 2019).

---

[6] A counterfactual possibility needs not even diverge from the actual event. One can mentally rewind the tape of what happened, replay the tape while re-rolling the die, and end up in the same situation. In fact our theory implies that many of the counterfactual possibilities that people simulate are identical to what actually happened.

Lucas and Kemp (2015, 2012) developed a formal model of counterfactual reasoning that accurately predicts how people answer questions about counterfactual conditionals. Their Extended Structural Model (XSM) is an updated version (modified to be more consistent with human intuitions) of a foundational model of counterfactual reasoning developed by computer scientist Judea Pearl (2000, 2013). For our purposes we do not need to use all components of the model – we only need the part of the model that describes how people simulate counterfactual possibilities.

According to the XSM, the propensity with which we simulate possibilities is determined by a trade-off between prior probability and what happened in the actual world. This trade-off is controlled by a stability parameter s. When we simulate a counterfactual possibility, for each exogenous variable in the model, with probability s we set the variable's value to the one it has in the actual world. With probability $1 - s$, we instead sample the variable's value from its prior probability distribution (Lucas & Kemp, 2015)[7].

Saying the same thing more formally, the sampling propensity of variable X is given by the equation:

$$SP(X = x) = s\delta(x) + (1 - s)Pr(x)$$

Where $SP(X = x)$ denotes the sampling propensity of $X = x$, and $\delta(x)$ is 1 if $X = x$ in the actual world, and 0 otherwise[8]. See Figure 2 for graphical illustration.

For example, when sampling the value of *Green* from our causal model of the game, with probability s we simply copy the value of *Green* in the actual world (i.e. 1); with probability $1 - s$ we instead sample it from its prior probability distribution (i.e. 0.1). For $s = .5$, this would mean that overall we sample $Green = 1$ with probability 0.55.

[7] In principle, some variables may be more or less stable than others. But to avoid making the model too flexible, we will use the same single value of s across all variables and all experiments we report here.

[8] More generally, we can assume that $\delta(x)$ is the posterior probability of $X = x$ in the actual world conditional on what we were able to observe about the actual world. In cases where we know the actual-world value of X, this is equivalent to the formulation above.

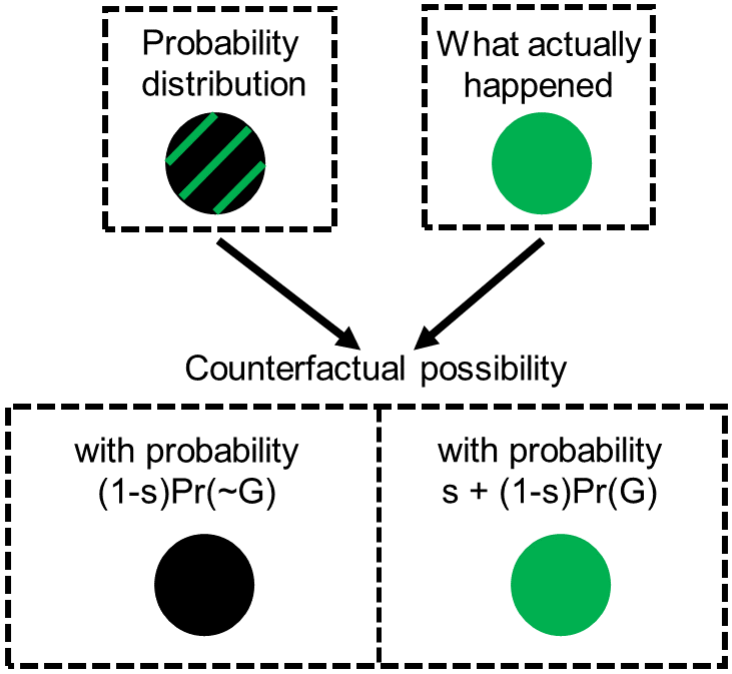

Figure 2

How people sample an exogenous variable, according to the Extended Structural Model of counterfactual reasoning (Lucas & Kemp, 2015). Here we illustrate how people determine whether Alice draws a green ball from the first box when they simulate a counterfactual possibility, in our game example. The model sometimes samples from the probability distribution over possible draws from the first box (with probability $1 - s$*), and sometimes simply copies what happened in the actual world (with probability* s*).*

After sampling the value of the exogenous variables, the value of the endogenous variables are then set accordingly, as a function of the causal relationships encoded in the causal model's functional equations.

The XSM accurately describes the way people answer questions about counterfactual conditionals across data from many experiments (Lucas & Kemp, 2015; Sloman & Lagnado, 2005; Rips, 2010). Notably, the model is a better fit to people's intuitions than alternative models that assume that people sample exogenous variables exclusively from their prior probability distributions, or that people disregard prior

probability altogether.

What are the implications of this work for causal judgment? Remember that according to our framework, when people make causal judgments they sample counterfactual possibilities from their causal model of the situation. We now make the hypothesis that the counterfactual possibilities that people use in their causal judgments are generated in the same way as the counterfactual possibilities people generate when they answer questions about counterfactual conditionals. The XSM provides a formal model of this process of counterfactual sampling.

The second question facing a counterfactual account of causal selection is the following. What computations do people apply to the counterfactuals they generate?

A formal theory of causal strength

Following a recent proposal (Quillien, 2020), we suggest that people make causal judgments by computing a statistical measure of 'effect size' across counterfactuals. However, an alternative hypothesis is that people make causal judgments by thinking about abstract features of the relationship between the candidate cause and the outcome, such as whether the candidate cause was necessary and sufficient for the outcome. Icard and colleagues (2017) have developed a computational model formalizing this latter hypothesis.

We focus on these two theories because (a) they fit naturally within the counterfactual framework we use here, (b) they make quantitative predictions, and (c) they have been particularly successful at predicting people's causal judgments across a wide range of tasks (Quillien & Barlev, 2022; Henne, Kulesza, et al., 2021; Kirfel et al., 2021; Morris et al., 2019; O'Neill et al., 2021; Gerstenberg & Icard, 2020; Henne et al., 2019; Gill et al., 2022). We discuss other theories of causal judgment in the General Discussion.

The two theories we focus on share the basic skeleton we outlined earlier. That is, they assume that people make causal judgments by sampling counterfactuals from their causal model of the world, and then use these counterfactuals as an input to some

computation. They differ in the form that this computation is supposed to take.

Counterfactual Effect Size Model (CESM)

According to the Counterfactual Effect Size Model (CESM; Quillien, 2020), the causal strength of C for E quantifies how much an intervention on C would change E on average, across a variety of possible background circumstances. We first give an intuition for how the model works in the general case (and formally define the model in the Supplementary Information), then describe the very simple form it takes in the causal structures used in our experiments.

To understand how the CESM works, consider first a very simple way that one might define the causal strength of C for E. This simple way consists in imagining making an intervention on C while holding everything else about the situation constant, and computing the extent to which this intervention would change the value of E. For example, prevent the lightning bolt from striking the tree, and compute whether the forest still catches on fire. The extent to which E changes in response to an intervention on C is the causal effect of C for E. While somewhat intuitive, this definition neglects the intuitive requirement for causal judgment to identify causes that would have led to the outcome even if background circumstances had been somewhat different.

The CESM retains the basic intuition behind the simple definition, but adds the idea that we should repeatedly perform the causal effect computation while varying the background circumstances. As such, the model repeats the following process a large number of times: Simulate a counterfactual possibility by re-sampling all exogenous variables (and re-computing the value of endogenous variables accordingly); then, compute the extent to which an intervention on C would result in a change in the value of E under the particular circumstances we have just sampled.

Repeating this process many times allows us to compute the average causal effect of C for E across various background circumstances. Importantly, we express this average as a

standardized effect size, by multiplying it by the ratio of the standard deviations of C and E across all simulations. See the Supplementary Information for a formal definition of the model.

In our experiments, for simplicity we will use causal structures where there is no "confounding" between variables.[9] In this kind of causal structure, the CESM has a very simple interpretation. The computations described above can be shown to be equivalent to simulating counterfactual possibilities, and computing the correlation between C and E across these counterfactuals (see Quillien, 2020 for proof).

Consider Alice, who drew a green ball from the first box and a blue ball from the second box, and (since she got two colored balls) won the game as a result. Did drawing a green ball cause her to win the game? The CESM holds that when people make this judgment, they compute the correlation, across the counterfactual possibilities that come to mind, between the events “Alice draws a Green ball” and “Alice wins the game”. People think that drawing a green ball caused Alice to win the game if this correlation is high (see Figure 3).

Necessity-Sufficiency Model (NSM)

The Necessity-Sufficiency model (Icard et al., 2017) is one formalization of the idea that causal judgment involves two kinds of counterfactual simulations. It holds that when people consider whether C caused E, they engage in retrospective simulation, asking whether C was necessary for E in that particular situation, and they also engage in prospective simulation, asking whether C is in general sufficient for E to happen. More formally, the model assumes that people compute two distinct measures of how C causally

[9] Technically, these are causal structures where C can have a causal influence on E, E cannot have a causal influence on C, and there is no variable that can have a causal influence on both C and E. That is, the no-confounding condition holds whenever $Pr(E = e|C = c) = Pr(E = e|do(C = c))$: the probability that E takes value e, given that we observe C take value c, is equal to the probability that $E = e$ given that we make an intervention setting C to c (see Pearl, 2000).

affects E: a measure of Necessity and a measure of Sufficiency. Then they compute the causal strength $K(C \rightarrow E)$ of C for E by computing a weighted mean of these two measures:

$$K(C \rightarrow E) = (1 - SP(C = 1))Necessity(C \rightarrow E) + SP(C = 1)Sufficiency(C \rightarrow E)$$

Where $SP(C = 1)$ is the sampling propensity of C, i.e. the probability of sampling $C = 1$. Icard et al. (2017) are not strongly committed to particular operationalizations of Necessity and Sufficiency, but they have suggested the following definitions.

Necessity. $Necessity(C \rightarrow E) = 1$ if, in the actual world, an intervention setting C to 0 would prevent E from happening.

Sufficiency. The Sufficiency strength of C for E is the probability that, in a situation where E and C do not happen, an intervention setting C to 1 would be enough to make E happen[10]. Formally:

$$Sufficiency(C \rightarrow E) = Pr(E = 1|do(C = 1), E = 0, C = 0)$$

It is worth clarifying what this quantity describes, in the context of the model of counterfactual sampling we have introduced in the previous section. The sufficiency score can be seen as the value we would reach (in the limit of an infinity of samples) if we did the following:

- sample repeatedly from our causal model of the situation,
- keep only those simulations where C = 0 and E = 0,
- compute the proportion of these simulations where an intervention setting C = 1 results in E = 1.

[10] Icard et al. (2017) also suggest that sufficiency might be defined as $Pr(E = 1|do(C = 1))$. In practice, we find that using this alternative definition does not improve the fit of the model to the data from the present experiments.

We will use the definitions of Necessity and Sufficiency we have just described in our implementation of the model.

Summary and overview of empirical tests

To recapitulate, we defend a theory of causal selection according to which, when people make a causal judgment about whether C caused E, they do the following (see Figure 3):

(i) They simulate counterfactual possibilities for what could have happened, by sampling them from their causal model of the situation. They tend to sample possibilities that are both likely and close to what actually happened (Lucas & Kemp, 2015).

(ii) They compute a measure of effect size that quantifies the average effect that an intervention on C would have on E across these counterfactual possibilities. In simple situations, this is equivalent to computing the correlation between C and E across counterfactual possibilities (Quillien, 2020).

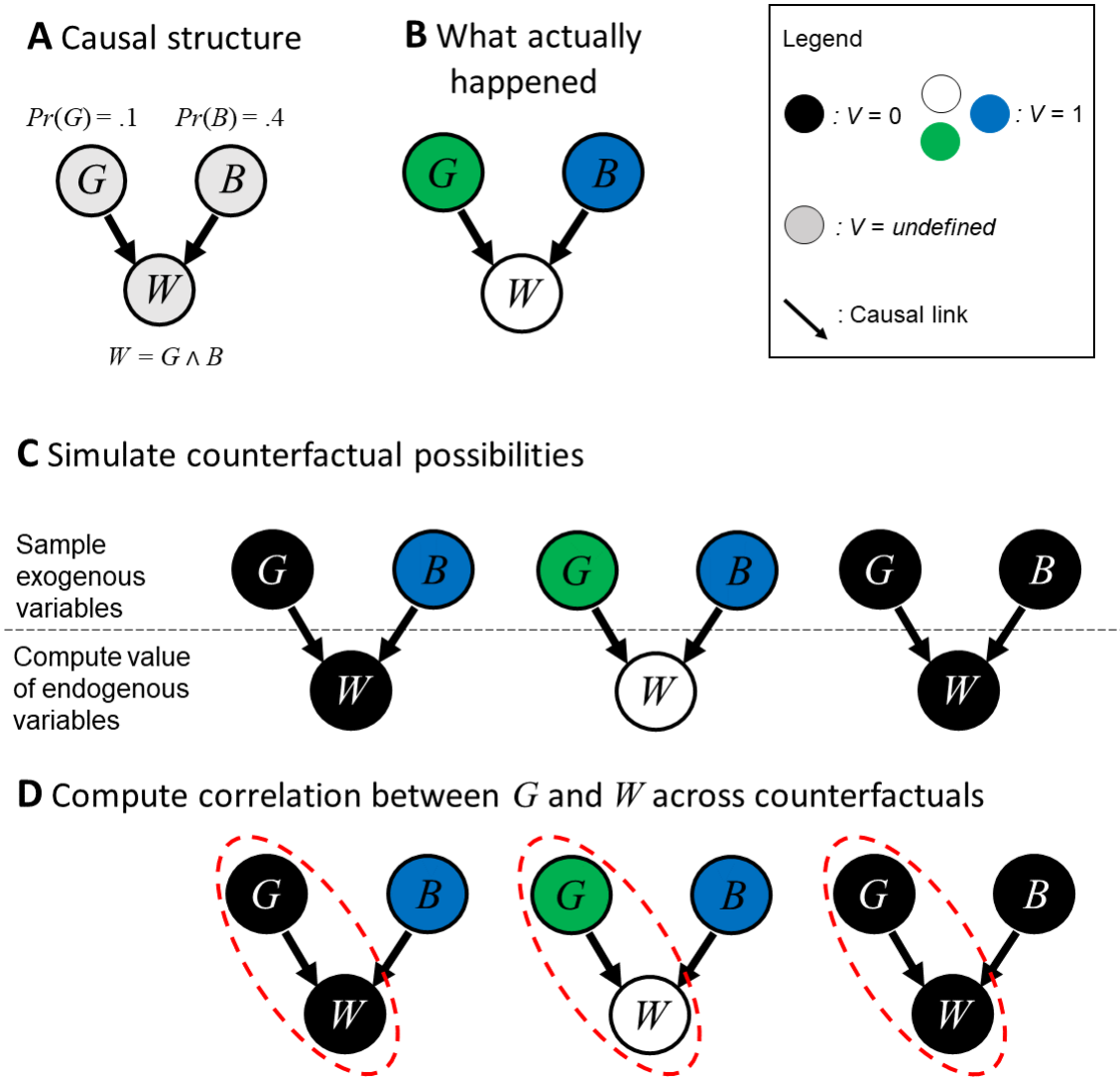

Figure 3

Schematic representation of our theory. *Here we model how people judge whether drawing a green ball from the first box caused Alice to win the game.* ***A)*** *The causal structure of the situation.* ***B)*** *In the actual world, Alice drew a green ball, a blue ball, and won the game.* ***C)*** *We simulate counterfactual possibilities that are biased toward situations that are a priori likely (we draw Blue more often than Green) and similar to what actually happened (we over-sample Blue and Green).* ***D)*** *Across counterfactuals, drawing a green ball is highly correlated with winning the game. Therefore we judge that Alice won the game because she drew a green ball.*

In what follows, we test whether this theory accounts for how people make causal judgments. First we re-analyze empirical data from existing studies. Then we conduct new experiments designed to discriminate between our account and some alternatives. Studies 1, 2 and 4 test divergent predictions of the NSM and CESM. Studies 2, 3, 4, test our assumption that counterfactual sampling is centered on what actually happened.

Reanalysis of existing data

We selected studies on causal judgment that had a sufficient number of observations to allow a meaningful measure of model fit. These studies collectively cover a wide range of manipulations, stimuli, and dependent variables (Quillien & Barlev, 2022; Lagnado et al., 2013; Morris et al., 2019; O'Neill et al., 2021; Zultan et al., 2012).

For each of the nine studies we selected, we computed model fit (following Morris et al., 2019; O'Neill et al., 2021) as the correlation between the CESM and average human judgments, across possible values of the stability parameter s. We also conducted the same analysis for the NSM. We describe the methodology in more detail in the Supplementary Information (https://osf.io/vnh84).

As Figure 4 shows, the CESM provides a good account of participants' causal judgments in all studies. In the Supplementary Information, we include for each study a graph depicting the human data alongside with the predictions made by the CESM for $s = .73$, which is the value of the stability parameter that results in the best fit in our new experiments.

First, the model is able to account for people's attributions of blame and responsibility in a series of studies by Zultan et al. (2012) and Lagnado et al. (2013). The authors asked participants to read about teams trying to achieve a goal whose completion depended in some way on the success of individual team members, and asked participants to what extent a given individual was blameworthy (in Zultan et al., 2012) or responsible (in Lagnado et al., 2013)for the failure or success of their team. The studies manipulated

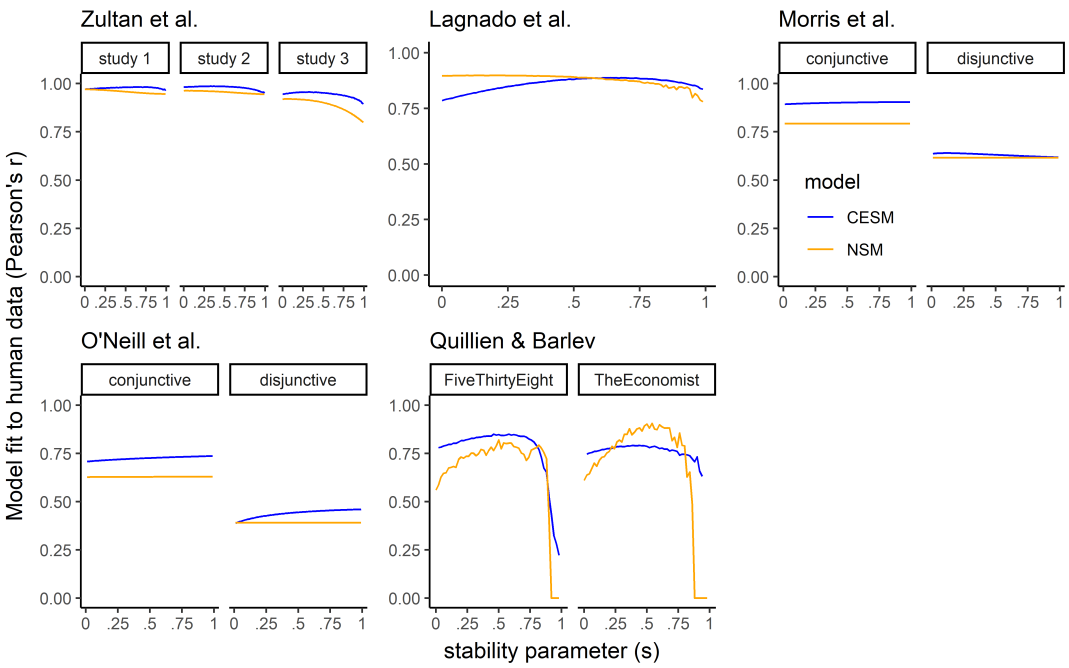

Figure 4

Correlation between model and mean human judgments as a function of stability parameter s, in large datasets from existing studies on causal judgment, for the CESM (blue) and the NSM (orange). The two panels for Quillien and Barlev (2022) correspond to the same human data, but with computational models calibrated with two different election forecasts (see original reference for details).

(i) the nature of the function that links individual successes to the team's success, and (ii) which team members failed or succeeded in a given trial.

Both manipulations influenced participant's attributions of responsibility. Lagnado et al. (2013) showed that these results could be explained by a computational model which assumes that people assign more responsibility to players who were *pivotal* for the outcome in this particular situation (i.e. they were necessary or close to being necessary, Chockler and Halpern, 2004) and who were *critical* for the outcome (they were a priori likely to be necessary; see Lagnado et al., 2013 for details). Using cross-validation, we find that the CESM has a slightly better fit to the data from Zultan et al. (2012) and Lagnado et al.

(2013) than the pivotality-criticality model[11].

Second, the CESM is able to account for the effects of manipulations of the prior probability of events, and also correctly predicts that the direction of the effect can reverse depending on the causal structure of the situation. Morris et al. (2019) and O'Neill et al. (2021) asked people to make causal judgments about an outcome which depended on two events. They manipulated the prior probability of the focal event (the event that participants were asked about) as well as the prior probability of the alternate event (the other event that contributed to the outcome) and the causal structure of the situation (in the conjunctive structure, both events were necessary for the outcome; in the disjunctive structure either event would have been sufficient). As shown in Figure 5, in conjunctive causal structures participants give high causal judgments for the focal event to the extent that (i) it was unlikely, (ii) the alternate was likely. In disjunctive causal structures, people have the opposite tendency: they give high causal judgments if the focal event was likely but the alternate was unlikely (Figure 6; see also Figures S2 and S4). This pattern of judgments is inconsistent with most theories of causal judgment (see discussion in Morris

[11] We evaluated the performance of each model by performing leave-one-out cross-validation, separately for each of the three studies in Zultan et al. (2012) and for the study in (Lagnado et al., 2013). That is, for each study we repeatedly divided the data into a training set (which excluded the data from one condition) and a test set (which consisted in the data from the condition excluded from the training test), fit the model's free parameter(s) on the training test and computed the error in the test set. We find that in each study the CESM has a slightly lower Root Mean Squared Error (RMSE) than the pivotality-criticality model, indicating a slightly better fit to the data (Zultan et al., 2012, Study 1: RMSE = .23 vs .25, Study 2: RMSE = .17 vs .25; Study 3: RMSE = .33 vs .35; Lagnado et al., 2013, RMSE = .48 vs .51). We did not evaluate the performance of the pivotality-criticality model on the other datasets, because it was not explicitly designed as a general theory of causal attributions beyond judgments of blame and responsibility. The model also fails to predict that causal judgments about an event are influenced by the event's prior probability. In the Supplementary Information we explore the relationship between the CESM and pivotality / criticality in more detail.

et al., 2019), but is predicted by the CESM.[12]

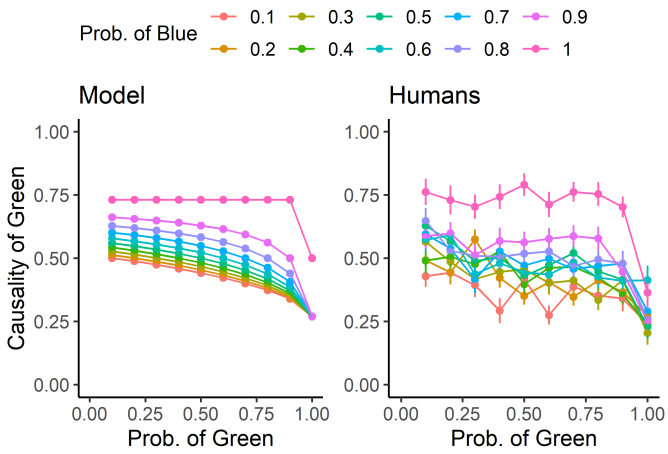

Figure 5

Human data and CESM judgments (for $s = .73$) for experiment 1 in Morris et al., 2019. Human judgments are standardized on the [0,1] interval.

Third, the CESM accounts for people's judgments about a complex real-world event. Quillien and Barlev (2022) asked participants how much each state that Joe Biden won in the 2020 presidential election caused him to win the presidency. We find that the model predicts participants' intuitions for a wide range of values of the stability parameter s.

On the other hand, Figure 4 shows that existing data do not completely rule out alternative hypotheses about how people make causal judgments. First, people might just be sampling counterfactual possibilities from their prior probability distributions. That is, when setting the stability parameter s to 0, our theory effectively ignores what happened in the actual world, but still gives a reasonable account of people's causal judgments in existing data. We will refer to this special case of our model as the "Unattached CESM". Second, the Necessity-Sufficiency model (NSM) also has a good fit to the data.

Therefore, we designed new experiments to arbitrate between our theory (which combines the CESM and the XSM) and these alternatives.

[12] This result was first reported in Quillien, 2020 for the special case of $s = 0$. Here we find that the pattern holds even when assuming that counterfactual sampling is biased toward what actually happened.

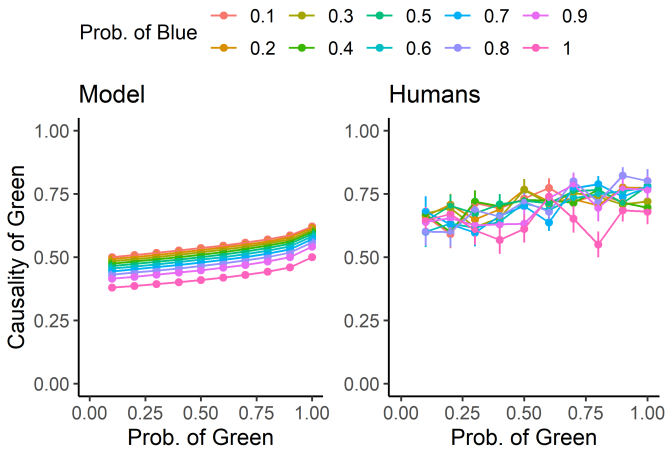

Figure 6

Human data and CESM judgments (for $s = .73$) for experiment 2 in Morris et al., 2019. Human judgments are standardized on the [0,1] interval.

Most existing studies on human causal judgment ask participants to reason about relatively simple causal structures, where two different causes lead to an outcome. In order to arbitrate between the alternative hypotheses we test here, we had to use slightly more complicated causal structures, with three or more potential causal variables. Using these more complicated settings raises the challenge of making sure that people manage to correctly represent the causal structure of the situation. Therefore, instead of asking participants to read simple vignettes, we developed an interactive task allowing participants to familiarize themselves with the setting (studies 1-3), or asked them to view video clips (Study 4). Data and R code for all studies are available at https://osf.io/h42f7/.

Computational Modeling

For each study we report below, we generated causal judgments for the CESM and the NSM, using the XSM as a process of counterfactual sampling. We asked participants to make judgments about the outcome of a simple game of chance. For each study we generated predictions for the CESM by simulating 500,000 possible rounds of the game according to the rules of that game, what happened in the actual situation described to

participants, and the sampling model described by the XSM.

We computed the CESM judgment for an event as the correlation between this event (for instance, whether the player makes a successful draw from an urn) and the outcome of the game (whether the player wins the game), across simulations. We computed NSM judgments analytically (see Supplementary Information), except for Study 1 where we computed sufficiency scores via simulation.

We fit the value of the stability parameter s by finding the value of s that results in the best average fit between model judgments and average participant judgments across studies 2, 3 and 4 (Study 1 does not have enough conditions to compute a meaningful measure of model fit). We quantified model fit using correlations between model judgments and average participant judgments (following Quillien and Barlev, 2022; Gerstenberg et al., 2021; Morris et al., 2019; O'Neill et al., 2021)[13]. For the CESM, the best-fitting value of s was $s = .73$ (i.e. 73% of the time we copy the value of an exogenous variable from its actual-world value)[14]. For the NSM we find $s = .15$. We use these values to generate CESM and NSM judgments in all the model-based analyses we report in Studies 2-4.

Study 1

One of the most basic findings about causal judgment is the *abnormal inflation* effect: in many contexts, people tend to select unexpected events as causes (Hart & Honoré, 1985; Morris et al., 2019; Hilton & Slugoski, 1986; Henne, O'Neill, et al., 2021). Surprisingly, the CESM and NSM offer very different explanations for why abnormal inflation should occur.

[13] We compute model fit within a given experiment, as we do not assume that the mapping between model judgments and participants judgments is necessarily identical across experiments. For example, a model judgment of .3 might result in more or less high ratings on a Likert scale depending on idiosyncratic features of the study stimuli.

[14] For comparison, Lucas & Kemp (2015) find a best-fitting value of $s = .53$ when analyzing their own experimental data, and $s = .77$ when analyzing experimental data from Rips (2010).

Remember that according to the NSM, the causal strength of C for E is a weighted sum of C's necessity and sufficiency, where the weights are the sampling propensity of C:

$$K(C \rightarrow E) = (1 - SP(C))Necessity(C \rightarrow E) + SP(C)Sufficiency(C \rightarrow E)$$

This formula says that the more unexpected an event, the more the event being necessary increases its causal strength. So, the NSM predicts that people select unlikely causes if these causes were necessary for the outcome.

By contrast, if C was not necessary for E, then the formula reduces to:

$$K(C \rightarrow E) = SP(C) * Sufficiency(C \rightarrow E)$$

In such a case, the causal strength of C is clearly proportional to its prior probability. In other words, when C was not necessary for E we expect the reverse of an abnormal inflation effect: people should select high-probability events as causes.

By contrast, according to the CESM, people favor causes that are highly correlated with the outcome, across counterfactuals. The model predicts that abnormal inflation will happen in situations where unlikely events tend to be correlated with the outcome, across counterfactuals.

Most empirical studies of causal judgments have focused on simple causal structures, where both models make the exact same predictions about the conditions in which abnormal inflation will occur (Icard et al., 2017; Morris et al., 2019; Gerstenberg & Icard, 2020; Henne, O'Neill, et al., 2021; Kominsky & Phillips, 2019). In Study 1, we design a situation where the models' predictions come apart. In our setting, a low-probability event is (i) highly correlated with the outcome across counterfactuals, but (ii) not necessary for the outcome. The CESM, but not the NSM, predicts an abnormal inflation effect. Specifically, we consider a simple situation where four events (two low-probability events and two high-probability events) lead to an outcome. But any combination of three events among these four would have been sufficient for the outcome to occur. Thus, no event was

individually necessary for the outcome. Yet, across counterfactuals, the outcome is more highly correlated with low-probability than high-probability events.

Method

We designed a simple game, which we made participants play so that they could get familiar with its rules. Then we instructed participants to consider the outcome of a round of the game played by someone else, and asked them about the causal contribution of each event.

Our game is a simple game of chance: the player has to randomly draw a ball from each of four urns, containing ten balls each. Each urn contains a mix of colored balls and black balls. A black ball gives 0 points, and a colored ball gives 1 point. The player wins if he or she gets 3 points or more after having drawn once from each urn.

For our main dependent variable, we showed participants the outcome of a game played by a fictitious other player, who drew a colored ball from each urn, and therefore won the game (see Figure 7). For each urn from which that player drew a colored ball, we asked participants to what extent they agreed that the player won because he drew a colored ball from that urn, on a Likert scale ranging from 1 (strongly disagree) to 9 (strongly agree).

To familiarize people with the game, we first asked them to play ten rounds. In each round of the game, four urns were displayed on the webpage, along with the participant's current score, and a reminder of the number of points needed to win the game. Participants could draw from an urn by clicking on a button next to that urn. Clicking the button made a ball (the outcome of that draw) appear next to the urn. Participants could draw from the urns in the order they wanted, but could not draw more than once from each urn in a given round. The outcomes of the draws were pseudo-randomly generated, with the constraint that the total proportion of colored balls that a participant would draw from a given urn, across the ten rounds, would exactly match the proportion of colored

balls in that urn (so that, e.g. if there are four colored balls in an urn, the player would draw a colored ball from that urn in exactly four out of the ten rounds of the game).

Urns 1 and 2 contained 1 colored ball each, while urns 3 and 4 contained 9 colored balls each (remember that each urn contained 10 balls in total). Thus urns 1 and 2 are "low-probability urns", while urns 3 and 4 are "high-probability urns" (these numerical and verbal labels were not shown to participants). The condition for winning the game was to score at least 3 points (i.e. drawing at least 3 colored balls out of four).

Across participants, the location of urns on the screen, and the positions of balls within an urn, were randomized. However, for a given participant, the urns and balls kept their location on the screen throughout the game, and this location was the same for the game they played and the game they witnessed. The urns were identified by letters (A through D) according to their location on the screen.

In the round of the game that participants witnessed, the fictitious player drew a colored ball from all urns, and therefore got 4 points, and won the game. Note that, because 3 points would have been enough to win the game, none of the successful draws were individually necessary to win the game.

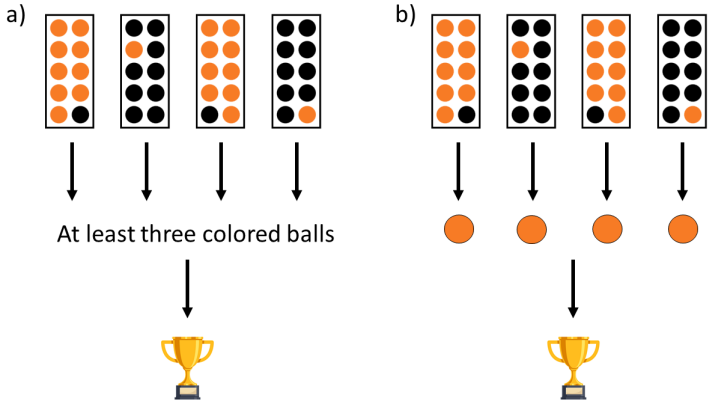

Figure 7

Causal structure and observed event, Study 1. a) *Rules of the game.* ***b)*** *Outcome of the game that participants make a causal judgment about. (The particular positions of balls and urns were randomized across participants.)*

Procedure

After completing a consent form, participants read a few pages of instructions about how the game works. Then they answered two multiple-choice comprehension questions: "How many points do you get when you draw a black ball?" (correct answer: 0), and "What is the condition for winning the game?" (correct answer: "you must get at least three points"). In the training phase, they played ten rounds of the game. In each round, their current score was displayed on top of the screen, and a reminder that they needed at least 3 points to win was displayed at the bottom of the screen.

After this phase, participants rated how difficult they thought it was to win the game, on a 1-9 Likert scale. In the test phase, participants were shown the outcome of a round of the game played by a fictitious other player, and rated the causal strength of each successful draw. Each question was displayed on a separate page. Each page displayed every urn (along with the outcome of the draw from each urn), as well as the fictitious player's score and the number of points required for winning. The urn that the question was about was highlighted with a green border. For each urn, we asked participants how much they agreed that the player won because he drew a colored ball from that urn, on a Likert scale ranging from 1 (strongly disagree) to 9 (strongly agree). Then participants answered a few demographic questions and were redirected to Prolific for payment.

All the current studies were approved by the Institutional Review Board at the University of Edinburgh, protocol number 2019/58792.

Modeling

We generated model predictions for the CESM and the NSM by running simulations, for a wide range of different values of the stability parameter s. We find that across all values of s, the CESM predicts abnormal inflation (a preference to view low-probability events as causal), while the NSM predicts abnormal deflation (i.e., a preference for high-probability events).

Participants

We recruited 40 US residents (28 female, 10 male, 2 other; mean age = 24, SD=10) from Prolific. We excluded from analysis participants who failed to correctly answer either of two comprehension questions, yielding a final sample of 37 participants. Participants were paid £0.75 for their participation.

Results

We find an abnormal inflation effect. Participants gave higher causal judgments to draws from low-probability urns (M=6.45, SD=2.61) than draws from high-probability urns (M=5.35, SD=2.68); see Figure 8. This result was supported by a multilevel regression with causal judgments as dependent variable and probability as predictor, with random intercepts for participants; b = -1.37, se(b)=.40, t=-3.39, p <.001.

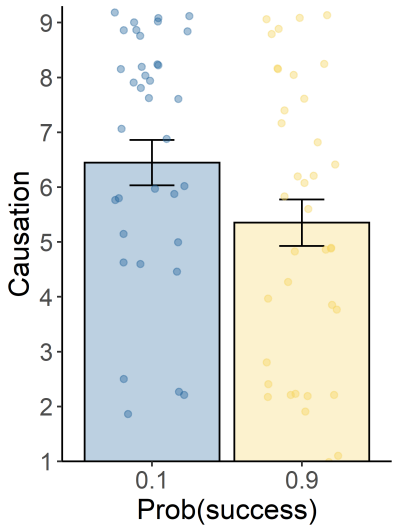

Figure 8

Mean human causal judgments for low-probability and high-probability urns, Study 1. Error bars represent the standard error of the mean. Each small dot represents the average causal judgments of one participant, across the two low-probability urns (in light blue), or the two high-probability urns (in light yellow).

Discussion

Study 1 suggests that abnormal inflation can be found even in cases where events are not individually necessary for the outcome, provided that low-probability events are highly correlated with the outcome, across counterfactuals. This is consistent with the CESM, but not with the NSM.

Might the current result simply follow from a general human bias to always select low-probability events as causes? Existing data strongly argue against this possibility: in many cases, people actually favor high-probability events as causes (Icard et al., 2017; Morris et al., 2019; Gerstenberg & Icard, 2020), or view high- and low-probability events as equally causal (Kirfel & Lagnado, 2021). Furthermore, in the next study we show that people sometimes give highest causal judgments to intermediate-probability events.

Study 2

Previous research has found in some settings, people's causal judgments are inversely proportional to the prior probability of the candidate cause, while in other settings, people select the most likely events as causal (see e.g. Morris et al., 2019). One possible interpretation of these results is that people are particularly drawn to events with extreme probabilities (either very high or very low). Yet according to the CESM there is nothing special in particular about extreme probabilities. In a setting where events with intermediate probabilities are most highly correlated with the outcome, the model predicts that people will select that event as the cause. We designed Study 2a to test this previously unexplored prediction.

We also designed an almost identical experiment, Study 2b, to test our assumptions about counterfactual sampling. Study 2b has the same causal structure as 2a (i.e. the rules of the game are the same), but what happens in the actual world differs. According to the XSM, people tend to sample counterfactuals that are likely, but also close to the actual world. Therefore, the model predicts that people should make different causal judgments in

the two studies.

Method

The design of Studies 2a and 2b was almost identical. We used a game similar to the one used in Study 1, except that there were three urns, each urn contained 20 balls, and the player needed 2 points or more to win the game. The low-probability urn contained 1 colored ball (probability of a successful draw : .05), the intermediate-probability urn contained 10 colored balls (probability : .5), and the high-probability urn contained 19 colored balls (probability : .95).

As in Study 1, participants first completed ten rounds of the game themselves, before observing the outcome of a game played by a fictitious player. The only difference between studies 2a and 2b was that in Study 2a, the fictitious player draws a colored ball from all three urns (thus getting 3 points), while in Study 2b, he draws a colored ball from the low-probability and the intermediate-probability urn, and draws a black ball from the high-probability urn (thus getting 2 points); see Figure 9.

The general procedure was identical to that in Study 1, except that for exploratory purposes we added three questions after the main study. These questions were designed to investigate whether participants have an explicit understanding of the way probability works in the urn scenarios used in the current setting. For each urn, we asked them how many times (on average) a player who randomly draws a ball from the box 20 times with replacement would get a colored ball. We pre-registered to classify as failing this task any participant who gave an answer different than 0 or 1 (for the low-probability urn), 9,10 or 11 (for the intermediate-probability urn), and 19 or 20 (for the high-probability urn), and to report analyses including these participants and other analyses excluding them. Interested readers can take the study at https://eco.ppls.ed.ac.uk/~tquillie/ql2022study2a/.

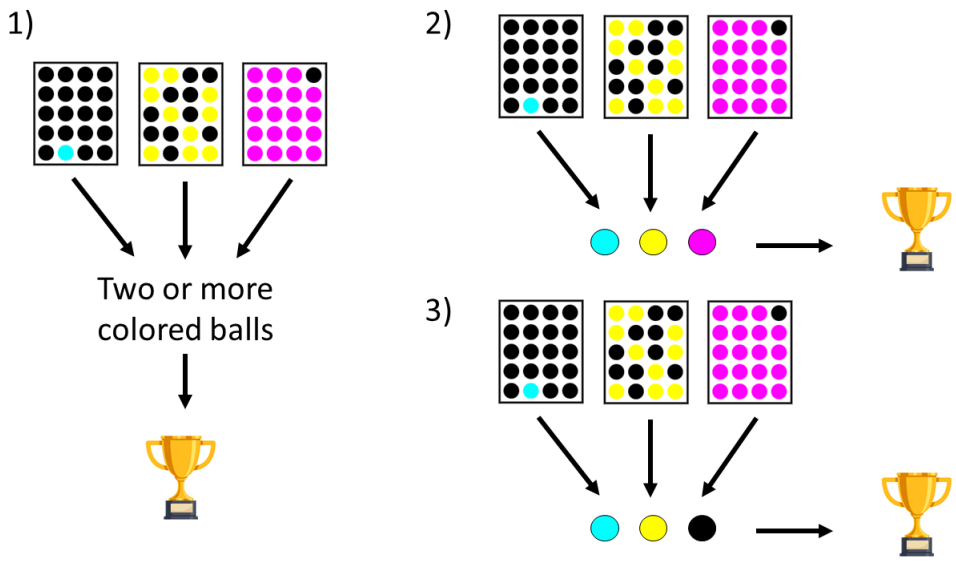

Figure 9

Causal structure and observed events, Study 2. ***1)*** *Rules of the game.* ***2)*** *Outcome of the game, study 2a.* ***3)*** *Outcome of the game, study 2b. (The particular positions and color assignments of balls and urns were randomized across participants.)*

Intuitive explanation of CESM predictions

The CESM predicts that, in Study 2a, people will judge the draw from the intermediate-probability urn as most causal. This is because in many of the counterfactual worlds sampled by the model, the player gets a black ball from the low-probability urn, and a colored ball from the high-probability urn. In such situations, the player wins the game if and only if he gets a colored ball from the intermediate-probability urn. Therefore, across counterfactuals, winning the game is most highly correlated with a successful draw from the intermediate-probability urn.

By contrast, in Study 2b, for high enough values of the stability parameter s, the CESM predicts that people will select the low-probability urn as most causal. This is because in the actual world, the player fails to draw a colored ball from the high-probability urn. Because the model is biased toward sampling counterfactuals that are similar to the actual world, it generates many counterfactuals where the player draws a

black ball from the high-probability urn. In such counterfactuals, the player needs to draw a colored ball from both other urns (i.e. both the low- and intermediate-probability urns) in order to win the game. This means that the low-probability urn becomes the limiting factor on the player's ability to win the game. As a result, across counterfactuals, winning the game is most highly correlated with a successful draw from the low-probability urn.

Participants

For Study 2a, we recruited 290 US residents from Prolific. We excluded from analysis 15 participants who failed either one of two comprehension questions (these were the same questions as in Study 1), yielding a total of 275 participants (148 female, 122 male, 5 other; mean age = 30.05, sd=2.92). Eighty-four of these participants failed at least one of our exploratory probability comprehension questions – following our pre-registered analysis plan we will conduct two analyses, first excluding and then including these participants.

For Study 2b, we recruited 290 US residents from Prolific. We excluded from analysis 23 participants who failed one or more comprehension questions, yielding a final sample of 267 participants (152 female, 110 male, 5 other; mean age = 33.6, sd age = 5.3). Seventy of these participants failed at least one of our exploratory probability comprehension questions – following our pre-registered analysis plan we will conduct two analyses, first excluding and then including these participants.

For both studies, participation was restricted to users who had joined the platform before 07/23/2021, because the platform had reported data quality issues due to an influx of users at that date[15]. Both studies were pre-registered (see https://osf.io/6cg4r for Study 2a and https://osf.io/qv5g4 for Study 2b).

[15] https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/

Results

Study 2a

As predicted, participants judged the successful draw from the intermediate-probability urn as most causal, see Figure 10. A repeated-measures Anova revealed a significant effect of probability on causal judgments, $F(2,380) = 15.61$, $p < .001$. Participants gave higher judgments to the intermediate-probability urn (M=6.52, SD=2.22) compared to the low-probability urn (M=5.52, SD=3.01), $t(190) = 3.84$, $p < .001$, and compared to the high-probability urn (M = 5.12, SD = 2.95), $t(190) = 6.98$, $p < .001$. Causal judgments were descriptively higher for the low- compared to high-probability urn, but this difference was not statistically significant, $t(190) = 1.34$, $p = .18$.

We replicated this analysis while including the participants who failed at least one exploratory probability comprehension questions. We find that results are essentially similar. A repeated-measures ANOVA found a significant effect of probability on causal judgment, $F(2, 548) = 21.38$, $p < .001$. Judgments for the intermediate-probability urn (M=6.57, SD=2.16) were higher than for the low-probability urn (M=5.67, SD=2.96), $t(274) = -4.17$, $p < .001$, and higher than for the high-probability urn (M=5.20, SD=2.93), $t(274) = 8.41$, $p < .001$. Although judgments for the low-probability urn were higher than judgments for the high-probability urn, this difference was not statistically significant, $t(274) = 1.86$, $p = .06$.

Study 2b

Participants considered the low-probability urn as more causal (M=7.65, SD=1.85) than the intermediate-probability urn (M=6.36, SD=2.22), $t(196)=6.2$, $p < .001$. We also conducted the same analysis while including an additional 70 participants who failed at least one probability comprehension question. This yielded a similar result: participants considered the low-probability urn more causal (M=7.44, SD=1.99) than the intermediate-probability urn (M=6.42, SD=2.20), $t(266)=5.6$, $p <.001$.

Note that this is the reverse of the effect observed in Study 2a (where the intermediate-probability urn was judged to more causal than the low-probability urn) – a reversal predicted by our theory. A 2*2 mixed Anova computed over the data from both studies (excluding judgments about the high-probability urn) confirms that this reversal is statistically significant; interaction effect, $F(1, 540) = 49.0$, $p < .001$.

As we see in the next section, this pattern of effects is predicted by the CESM, assuming that people sample counterfactuals according to the XSM.

Model-based analysis

The causal judgments made by each model, along with human judgments, are shown in Figure 10. CESM and average human judgments were correlated at $r(3) = .99$, $p < .001$, while NSM and human judgments were correlated at $r(3) = .67$, $p = .21$, and judgments from the Unattached CESM were correlated with human judgments at $r(3) = .25$, $p = .69$.

We also computed model fit on the non-aggregated data, using the Bayesian Information Criterion (BIC). Again, the CESM had a better fit to the data (BIC=5698.6), than the NSM (BIC=5747.5) and the Unattached CESM (BIC=5993.7).[16]

Discussion

Study 2a provides evidence for a previously untested prediction of the CESM: in settings where the outcome is most highly correlated, across counterfactuals, with intermediate-probability events, people consider intermediate-probability events as most causal. In addition, studies 2a-b jointly provide evidence for the hypothesis that when they

[16] Lower BIC values indicate better fit. Computing the BIC requires making assumptions about the probability that a participant would make a given judgment, given a model prediction. Here we modeled participants' judgments as samples from a truncated-discretized normal distribution with mean m and variance σ^2, where m is derived from the model prediction x_m via the logistic function $f(x) = \frac{8}{1+exp(-k(x_m - x_0))} + 1$. We fit the parameters k, x_0 and σ at the group level. We use the same approach to computing BIC in studies 3 and 4.

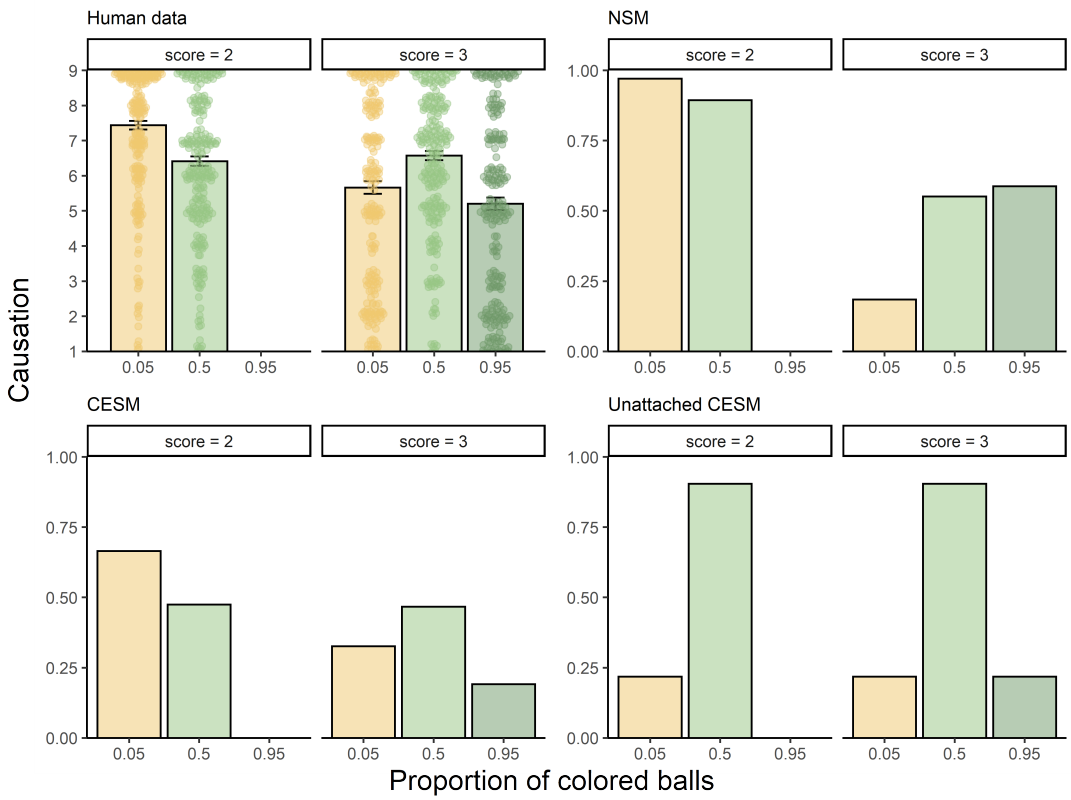

Figure 10

Human data (upper left), along with predictions of the CESM (lower left), NSM (upper right), and Unattached CESM (lower right) across Studies 2a-b. Studies are identified by the score of the observed player (3 points for Study 2a, 2 points for Study 2b). Error bars represent the standard error of the mean.

make causal judgments, people sample counterfactuals that are both likely and close to the actual world, in the manner predicted by the XSM. The only difference between studies 2a and 2b was what had happened in the actual world, and this difference was enough to reverse the effect of prior probability on causal judgments across the two studies. This is evidence that people's counterfactual samples are 'centered' on what happened in the actual world.

Study 2b also functions as a 'control' condition which rules out one possible interpretation of the results of Study 2a. In Study 2a, we asked participants to play ten rounds of the game themselves, to get familiar with the game, before asking them to make

causal judgments. Thus they already knew, before having to make causal judgments, that the game's outcome was most highly correlated with a successful draw from the high-probability urn. Could they have simply used this information to guide their causal judgments, without engaging in counterfactual reasoning? Study 2b rules out this interpretation. If people's causal judgments were simply guided by learned associations during the training phase, they would make similar judgments in studies 2a and 2b. That is, they would pick the intermediate-probability urn as the cause in Study 2b, because the training phase for that study was identical with that for Study 2a. But in fact they selected the draw from the low-probability urn in Study 2b, just like our counterfactual theory predicts.

In contrast to the CESM, the NSM was not able to fully reproduce the current pattern of results. In particular, the NSM predicts that in Study 2a, people should judge the draw from the high-probability urn as the main cause of the outcome. People judged the draw from that urn as being the least causal.

Study 3

The design of Study 3 is very similar to Study 2, except that it provides an even sharper test of our assumption about how people sample counterfactual possibilities. We used a variant of our game where participants draw from 3 urns, and two of the urns have (e.g.) purple colored balls, while the other urn has orange colored balls. To win the game, the player must draw at least one purple ball and one orange ball.

We chose the proportion of colored balls in each urn such that the outcome of the game is in general only moderately correlated with drawing an orange ball. But consider a situation where the player makes successful draws from all three urns (and so draws two purple balls and one orange ball). Since the player has drawn 2 purple balls, there are very few nearby counterfactuals where he draws 0 purple balls. If counterfactual sampling is biased toward nearby possible worlds, then participants should view the orange ball as

having been most causally important for winning the game, because in these counterfactuals it is almost guaranteed that the player has drawn at least one purple ball. By contrast, in a situation where the player gets only one purple ball (from an urn that has mostly black balls), participants should judge that drawing that purple ball was more causally important than drawing the orange ball.

Methods

We used the same setup as in Study 2, with the exception that two urns had purple balls, and the other urn had orange balls (we counterbalanced the color of the single urn across participants, but for ease of exposition we will pretend that the single-urn color is always orange). The player needs to draw at least one orange ball and one purple ball in order to win. There was one urn with 19 orange balls out of 20 (proportion of colored balls: 95%). The urns containing purple balls had 90% and 5% of colored balls, respectively. We will refer to them as the "Orange urn", "High-probability Purple urn" and "Low-probability Purple urn".

Across participants, we manipulated the balls that the fictitious player draws. In the "Two-events" condition, the player draws a colored ball from both the Orange urn and the Low-probability Purple urn, but draws a black ball from the High-probability Purple urn. In the "Three-events" condition, the player draws a colored ball from all three urns; see Figure 11. Note that the difference between these two conditions is very similar to the difference between studies 2a and 2b, but here we ran the two conditions at the same time, so we treat them as between-subject conditions in the same study rather than separate studies. The general procedure was identical to that in Study 2, except that we removed the three questions that looked at whether participants understand the way probability works. The study was pre-registered at https://osf.io/dfz5p/.

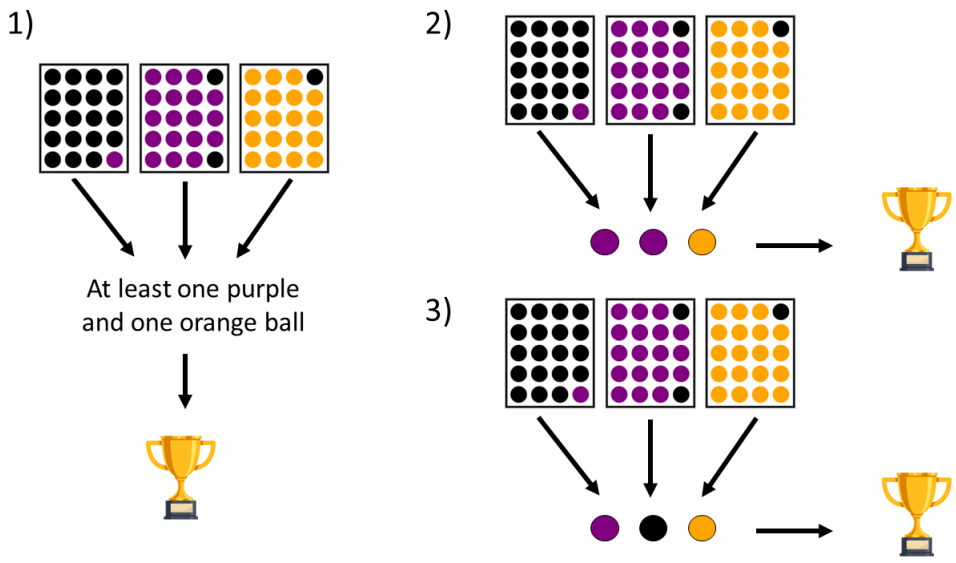

Figure 11

Causal structure and observed events, Study 3. ***1)*** *Rules of the game.* ***2)*** *Outcome of the game, Three-Event condition.* ***3)*** *Outcome of the game, Two-Event condition. (The particular positions and color assignments of balls and urns were randomized across participants.)*

Intuitive explanation of CESM predictions

Our main goal in Study 3 was to design a setting where the CESM and the Unattached CESM make maximally divergent predictions. The Unattached CESM is a special case of the CESM where s = 0: the model only samples counterfactuals from their prior probability distribution, without considering what happened in the actual world.

In both conditions, the CESM and the Unattached CESM make diverging predictions (except for trivially low values of s in the full CESM).

In the Three-Event conditions, the Unattached CESM predicts that the draw from the High-probability Purple urn is the most causal. Most of the time, the player wins the game by drawing a colored ball from the Orange urn and from the High-probability Purple urn, but the latter urn has fewer balls, and so it is usually the main bottleneck on whether the player wins the game. By contrast, the CESM judges that the draw from the Orange

urn is most causal. In possible worlds that are close to what actually happened, the player almost always draws at least one purple ball. Therefore whether the player wins mostly depends on whether he draws an orange ball.

In the Two-Event condition, the Unattached CESM predicts that the draw from the Orange urn is more causal than the draw from the Low-probability Purple urn. In general, the Low-probability Purple urn does not have much influence on the game's outcome, because the player rarely needs to draw a colored ball from it in order to win. By contrast the CESM predicts that the Low-probability Purple urn is most causal. In possible worlds that are close to what actually happened, the player rarely tends to draw a black ball from the High-probability Purple urn, and therefore the player needs to draw a colored ball from the Low-probability Purple urn to win the game.

Note that the CESM predicts the following cross-over interaction: in the Three-Event condition, the Orange urn is more causal than the Low-probability Purple urn, but the reverse is true in the Two-event condition.

Participants

We recruited 591 US residents (371 female, 211 male, 9 other; mean age = 34.2, sd = 12.5) from Prolific. Participation was restricted to users with a 90% or greater approval rate, who had taken between 50 and 1000 studies on the platform. We excluded from analysis 46 participants who failed at least one comprehension question, yielding a final sample of 545 participants.

Results

Three-Event condition

As predicted, the relative ranking of participants' causal judgments was Orange urn > High-Probability Purple urn > Low-probability Purple urn. This is consistent with the CESM but not with the Unattached CESM.

A repeated-measures Anova showed that participants' causal judgments significantly depended on which urn the question was about, $F(2, 580) = 181.8$, $p < .001$. Causal judgments for the draw from the Orange urn ($M = 7.44$, $SD=2.03$) were higher than for the High-probability Purple urn ($M = 6.17$, $SD = 2.46$), $t(290) = 9.18$, $p < .001$, and the Low-probability Purple urn ($M = 4.11$, $SD = 2.63$), $t(290)=18.47$, $p < .001$. Causal judgments for the High-probability Purple urn were higher than for the Low-probability Purple urn, $t(290) = 10.10$, $p < .001$.

Two-Event condition

As predicted, causal judgments for the Low-probability Purple urn ($M = 7.57$, $SD = 1.95$) were higher than for the Orange urn ($M = 6.95$, $SD = 2.31$), $t(253) = 3.0$, $p = .003$. Again, this is consistent with the CESM rather the Unattached CESM.

Note that this is the reverse of the relative ranking of the two urns in the Three-Event condition. A 2*2 mixed Anova, with condition and urn (Orange vs Low-probability Purple) as predictors confirmed that the cross-over interaction was significant, $F(1,543) = 275.5$, $p < .001$.

Model-based results

The correlation between the CESM and average human causal judgments was $r(3) = .80$, $p = .11$, while the correlation between Unattached CESM and average causal judgments was $r(3) = .30$, $p = .62$. The NSM had the best fit to the data, with a correlation of $r(3) = .98$, $p = .004$.

We also computed model fit on the non-aggregated data, using the Bayesian Information Criterion (BIC), and find the same relative ranking (NSM, BIC = 5508.8; CESM, BIC = 5509.2; Unattached CESM, BIC = 5757.4).

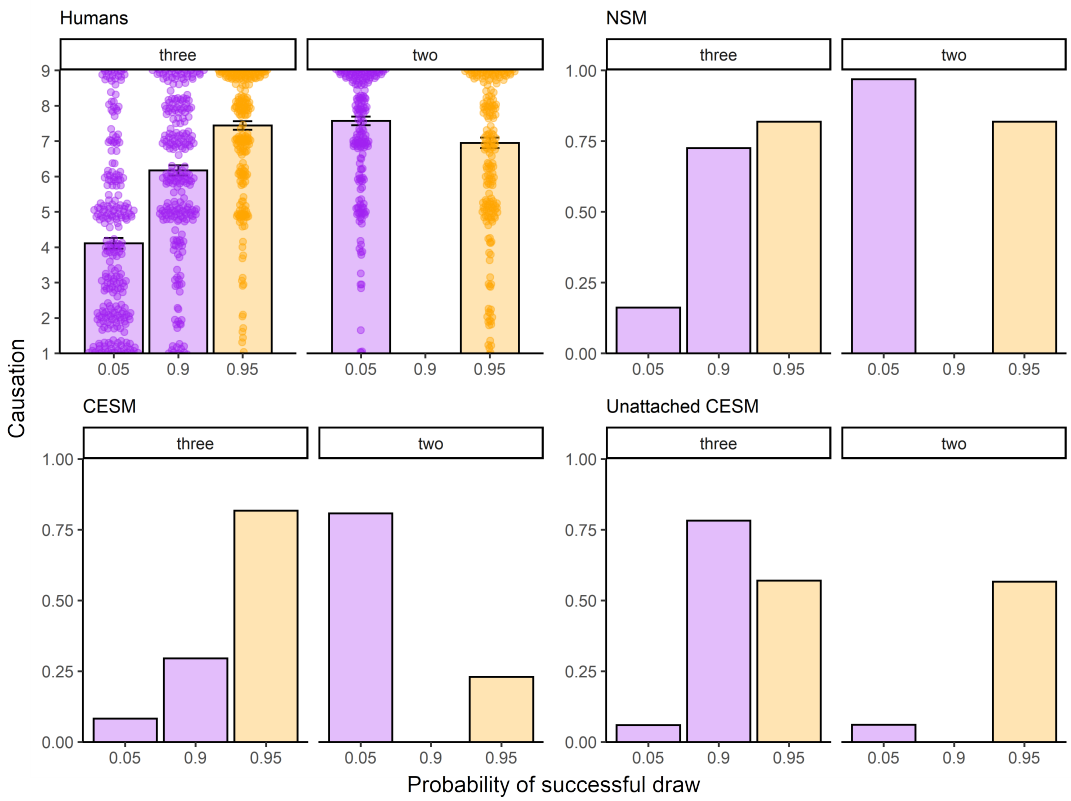

Figure 12

Human causal judgments, along with model predictions, Study 3. Error bars represent the standard error of the mean. The color contrast highlights the fact that two of the urns had colored balls of the same color as each other.

Discussion

In a setting where the CESM and Unattached CESM make sharply divergent predictions, we find that people's judgments are systematically aligned with the predictions made by the CESM with a non-zero value of the stability parameter s. In other words, people's causal judgments are most consistent with a model which assumes that people sample counterfactual possibilities that are both likely and close to the actual world, in the way specified by the XSM.

Study 3 was not designed to discriminate between the CESM and NSM, and both models make very similar predictions in this setting. Nonetheless, we find that the NSM

had a slightly better fit to people's causal judgments. While both models accurately predicted the relative ranking of causes within a given condition, the NSM additionally was able to predict the relative ranking of causes across conditions. While we find that the NSM fails to capture key features of our data in other studies, the model's good fit in Study 3 suggests that developing modified versions of the underlying theory might be an interesting avenue for future research.

Study 4

Study 4 aims to gather more evidence for our theory, using a new set of stimuli.

We designed another simple game of chance, inspired by the game of pinball. The game involves a simple sequence of events: three flippers randomly flip back and forth between two possible orientations before settling on a fixed orientation, and then a ball is released at the top of the screen. Depending on the top flipper's orientation, the ball is either directed to the left flipper or the right flipper. Then depending on that flipper's orientation, the ball either falls into a blue bucket or falls off the screen. The player wins if the ball ends up in a blue bucket. Flippers differ in their likelihood of flipping left or right – for example some flippers spend more time pointing toward the right than the left, and so their final orientation is more likely to be toward the right.

Imagine that the top flipper sent the ball to the right, and the right flipper then sent the ball toward the blue bucket (see figure 13). We are interested in how much people agree that the top flipper pointing to the right caused the player to win the game. Our theory makes the three following predictions (which we pre-registered; https://osf.io/q3xa4/).

Prediction 1

P1. If the left flipper pointed toward (as opposed to away from) a blue bucket, people will be less likely to think that the top flipper sending the ball to the right caused the player to win (see Figure 14).

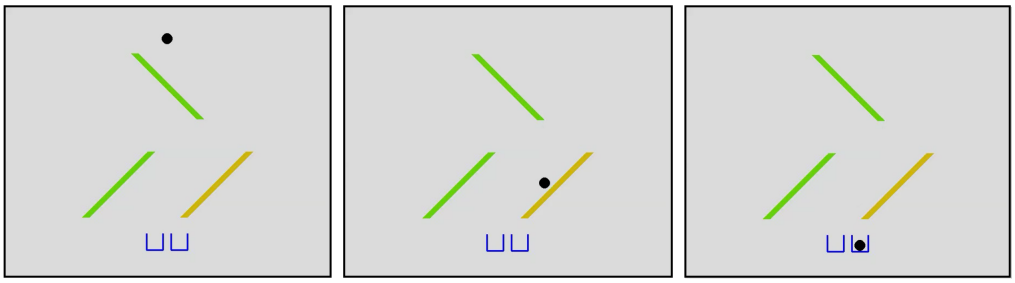

Figure 13

Frames from a video clip used in Study 4. The color of a flipper indicates its 'preferred' orientation.

Explanation. When the left flipper points toward a blue bucket, then the orientation of the top flipper does not really matter: the player will win the game either way. By contrast, when the left flipper points away from a blue bucket, the top flipper sending the ball to the right is crucial to the player winning the game. If people preferentially sample possible worlds that are similar to the actual world (as posited by the XSM), then in the first case they will tend to mostly sample possible worlds where the top flipper's orientation does not matter, and in the second case they will tend to mostly sample possible worlds where the orientation of the top flipper matters.

Prediction 2

Prediction 2 is more subtle, and follows from the combination of the XSM (specifically, the assumption that people preferentially sample counterfactuals that are similar to what actually happened) and the CESM:

P2. Consider a situation where the left flipper ended up pointing away from a blue bucket (figure 14 right panel). In that situation, there will be an abnormal inflation effect. That is, people will think that the top flipper played a large causal role to the extent that the top flipper pointing to the right is an unexpected event. But in a situation where the left flipper points towards a blue bucket (figure 14 left panel), there will be no abnormal

inflation effect.

Explanation. To understand this prediction, we need to understand what happens when we preferentially sample counterfactuals where the left flipper points away from the bucket. In such counterfactuals, the player wins the game if the top flipper sends the ball to the right AND the right flipper sends the ball toward a blue bucket. The outcome of the game thus depends on the state of the top and right flippers, and it will be especially correlated with the state of the top flipper to the extent that the top flipper has a low prior probability of pointing to the right. Therefore we predict an abnormal inflation effect if the left flipper points away from a blue bucket in the actual world.

By contrast, if we preferentially sample counterfactuals where both bottom flippers point toward a bucket, then there is no reason to expect an abnormal inflation effect, as the orientation of the top flipper will not be systematically correlated with the outcome.

Prediction 3

Prediction 3 follows from the CESM.

P3. If the left flipper has a general tendency to point away from a blue bucket, then people will give higher causal judgments to the top flipper sending the ball to the right.

Explanation. If the left Flipper Almost always points away from a blue bucket, then across possible rounds of the game the top flipper sending the ball to the right is correlated with the player winning the game. By contrast, if the left Flipper Almost always points toward a blue bucket then the top flipper sending the ball to the right is negatively correlated with the player winning the game.

In sum, the CESM predicts that the probability distribution over possible orientations for the left flipper will influence causal judgments. It is worth noting how this prediction differs from the effects of probability on causal judgments that have already been documented in the literature (e.g. Morris et al., 2019). Existing research has found that causal judgments about whether an event C caused an outcome depends on the prior

probability of C, as well as the prior probability of other causes that contributed to the outcome. Here we are interested in a factor that did not contribute to the outcome (since the ball did not reach the left flipper). Our prediction is that even probabilistic facts about that factor will influence causal judgments.

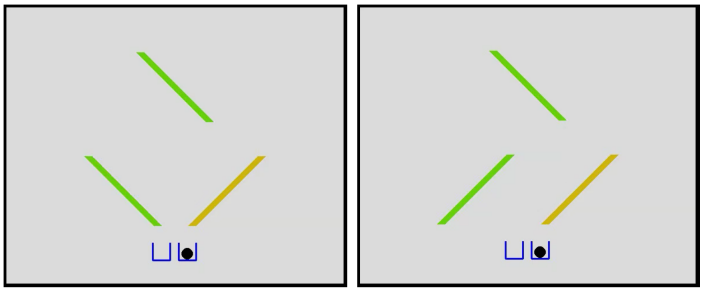

Figure 14

In some video clips, the left flipper (in its final orientation) points toward a bucket (left panel), while in other clips it points away from it (right panel).

Methods

We asked people to watch video clips depicting different rounds of the game where the player wins. At the end of each video clip we asked them to rate whether the top flipper sending the ball to the direction it did caused the player to win.

Materials

In each video clip, events proceeded as follows. First, the top flipper moved back and forth randomly between two orientations (left or right), before settling in one orientation. Then the same thing happened with the right flipper; then the same thing happened with the left flipper. After the left flipper had stopped moving, a ball was released from the top of the screen. The ball's trajectory then deterministically followed from the orientation of the flippers.

We manipulated the probability that a given flipper would orient in a particular direction. We told participants that some flippers had a "preferred orientation" which was

color-coded: green flippers tended to point toward the left, red flippers tended to point toward the right, and orange flippers were equally likely to point either way. When moving back and forth, a flipper spent 90% of the time in its preferred orientation (orange flippers spent 50% of the time in either orientation). At the side of each video there was a small box reminding participants of the color coding.

Design

For ease of exposition, we have focused on one example where the top flipper sent the ball to the right. In our experiment, we counterbalanced the direction the top flipper sent the ball, so in half of the trials, the flipper that the ball did not reach was the left flipper, and in the other half of trials it was the right flipper. We will refer to the top flipper as Flipper A, the flipper that was reached by the ball as Flipper B , and the flipper that was not reached by the ball as Flipper C.

We used a 2*2*2 mixed design. Between-subjects, we manipulated the actual-world orientation of Flipper C: in one condition ("C out" condition), Flipper C pointed away from a blue bucket, in the other condition ("C in" condition), Flipper C pointed toward a blue bucket (see figure 14). Within-subjects, we manipulated the prior probability of Flipper A sending the ball to where it sends it in the actual world (Pr(A)=.1 or .9), as well as the prior probability of Flipper C sending the ball toward a blue bucket (Pr(C) = .1 or .9). Pr(B) was always .5. For each of the four combinations of Pr(A) and Pr(C), we generated two video clips: one in which Flipper A sent the ball to the right, and one in which Flipper A sent the ball to the left. In the first kind of video clip, Flipper C was on the left, and on the second kind Flipper C was on the right[17]. Therefore, each participant watched 8 video clips.

In addition, we generated two video clips to serve as attention checks. These clips

[17] We counterbalance orientation such that Pr(A) and Pr(C) are not correlated with flipper color across trials.

had P(A) = .9, P(B)=.5, P(C)=.5, both bottom flippers pointed away from the goal in their final orientation, and the ball fell off screen. In one of them, the top flipper sent the ball to the left, and in the other the top flipper sent the ball to the right. After the ball fell off the screen, we masked the top flipper and asked participants whether the ball fell on the left or the right side of the screen. We excluded from analysis participants who answered at least one attention check question incorrectly.

Procedure

Participants completed a consent form, then read instructions about how the game worked. After completing two questions designed to check their understanding of the instructions, they proceeded to the task.

In the main phase, participants watched ten video clips (eight test videos and two attention check videos). After each test video clip finished playing, a question appeared on the screen, below the video, asking participants how much they agreed, on a 1-9 likert scale (from "strongly disagree" to "strongly agree") that "the player won the game because the top flipper sent the ball to [direction]", where [direction] was replaced with the final orientation of the top flipper (left or right) in that video.

Each video clip lasted 23 seconds, and participants were allowed to re-watch the video if they wanted (except on attention check trials), but they could not skip ahead on their first time watching the video. Video clips were presented in random order. Attention check videos were always presented on the third and seventh trials, and their order of presentation was counterbalanced.

Participants then answered a few demographic questions, and were redirected to Prolific for payment.

Participants

We recruited 389 US residents (209 female, 172 male, 8 other; mean age = 29.7, SD=8.7) from Prolific. Participation was restricted to users with a greater than 90%

approval rate, who had taken between 50 and 1000 previous studies. Recruitment and exclusion criteria were preregistered (https://osf.io/q3xa4). We excluded from analysis 32 participants who failed one or more comprehension questions (N= 15) and/or attention checks (N=22), for a final sample of 357 participants.

Results

Statistical tests reported below are linear mixed models, with participant-level random slopes and intercepts. Statistical significance for the effect of a given variable is assessed by an Anova comparison between the fit of a model with the predictor and the fit of an identical model omitting the variable.

We did not find any main effect or interaction effect involving whether Flipper B was the left- or right-side flipper, therefore we omit the variable from our analyses.

Figure 15 shows the human data, and the predictions of the computational models. We find support for all three of our pre-registered hypotheses:

H1: participants give higher causal judgments when the unreached flipper pointed away from the goal. Participants gave higher ratings in the Out condition (M=7.11, SD=1.92) than the In condition (M=4.95, SD=2.53), b=2.16, $p < .001$.

H2: Participants give higher causal judgments for lower values of Pr(A), but only in the "C out" condition. Participants in the "C out" condition gave higher causal judgments when Pr(A) was low (M=7.24, SD=1.82) than when Pr(A) was high (M=6.98, SD=2.01), b=-.33, $p < .01$. By contrast, participants in the "C in" condition gave identical causal judgments when Pr(A) was low (M=4.96, SD=2.52) as when it was high (M=4.94, SD=2.53), b = -.03, p = .73. The effect of Pr(A) on causal judgments was marginally higher in the "C out" than the "C in" condition, interaction effect: b=-.29, p=.05.

H3: participants give higher causal judgments for low values of Pr(C). Participants gave lower causal judgments when Pr(C) was high (M=5.91, SD=2.53) than

when Pr(C) was low (M=6.10, SD=5.91), b=-.22, $p < .001$.

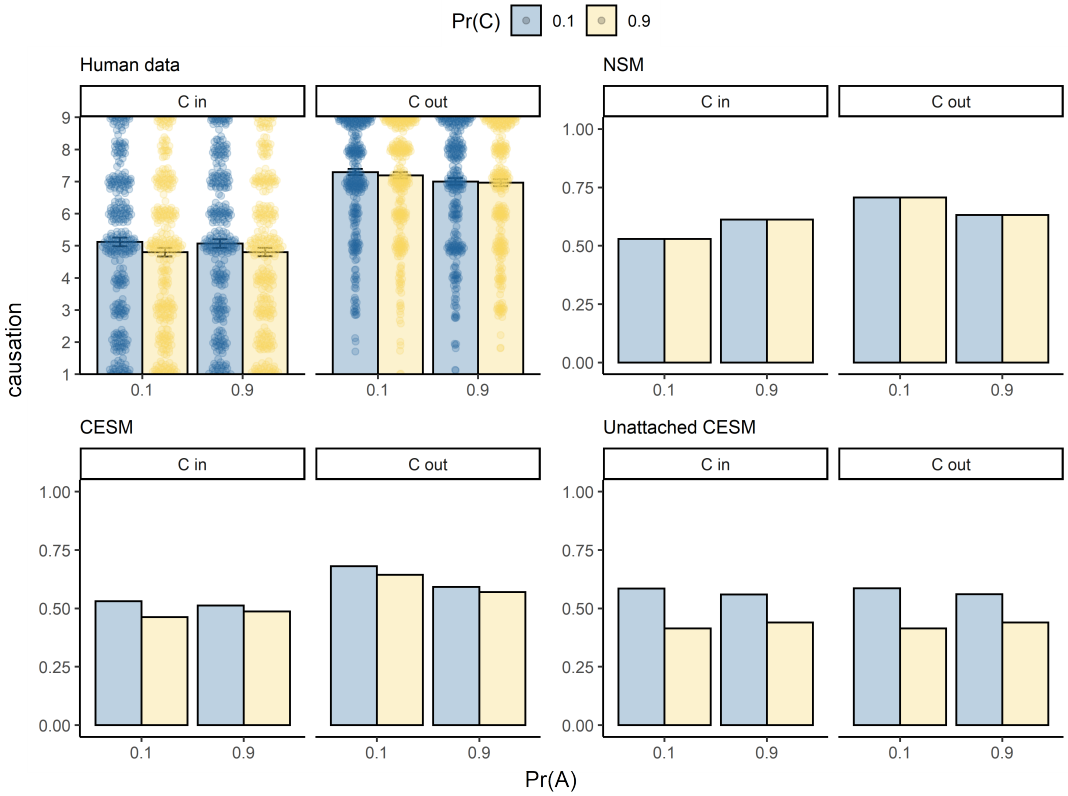

Figure 15

Human causal judgments, along with model predictions, Study 4. Error bars represent the standard error of the mean.

Model-based analysis

CESM judgments are negative for some of the conditions (especially for the Unattached CESM – for example if the right flipper is less likely than the left flipper to send the ball toward the blue bucket, then the top flipper sending the ball to the right is negatively correlated with the player winning), but since there is no meaningful demarcation between scores below or above 0 in the model, we pass all model judgments through a sigmoid function ($f(x) = \frac{e^x}{1+e^x}$) before plotting them. The statistical analyses that follow are performed with the untransformed model judgments.

CESM judgments were highly correlated with mean human judgments, $r(6) = .91$, p

= .002. NSM judgments were also correlated with mean human judgments, r(6) = .81, p = .02. By contrast, judgments by the Unattached CESM were not correlated with mean human judgments, r(6) = .08, p = .85.

We also computed model fit on the non-aggregated data, using the Bayesian Information Criterion (BIC), and find the same relative ranking (CESM, BIC = 11765.0; NSM, BIC = 11941.9; Unattached CESM, BIC = 12180.2).

On a qualitative level, the NSM could only partially reproduce the patterns in people's causal judgments. It was able to reproduce the higher causal judgments in the "C out" compared to the "C in" condition, and the abnormal inflation effect in the "C out" condition. However, it predicted an abnormal deflation effect in the "C in" condition that does not appear in the data, and failed to predict that Pr(C) influences causal judgments.

Discussion

Study 4 provides additional evidence that human causal judgment can be approximated by the CESM, assuming that people simulate counterfactuals in the way described by the Extended Structural Model (XSM) of counterfactual reasoning.

We find evidence for a subtle prediction of the CESM: people's causal judgments can be influenced by the prior probability of an event that did not contribute to the outcome.

Additionally, the data are consistent with the XSM's assumption that people sample counterfactuals that are both likely and close to the actual world. We find that people's judgments depended on what happened in the actual world in two ways. First, people gave higher causal judgments for an event when the outcome counterfactually depended on the event in the actual world. Second, by varying what happened in the actual world, we were able to modulate whether or not the probability of an event influenced participants' causal judgments.

Alternative counterfactual sampling models

Studies 2 to 4 provide evidence for the XSM's assumption that people sample counterfactual possibilities that are both likely and close to the actual world (Lucas & Kemp, 2015).

The following question now arises. Do our results simply depend on the general assumption that people sample counterfactual possibilities that are both likely and close to the actual world? Or do our results also depend on the specific hypothesis the XSM makes about how these two factors (probability and similarity to the actual world) are *integrated*?

To explore this question, we designed two alternative models of counterfactual sampling, and we generated causal judgments for the CESM under these models. Like the XSM, these models assume that sampling propensity is influenced by both probability and closeness to the actual world – but they make different assumptions about the details.

On the first model, for each counterfactual world people simulate, they either entirely copy that world from the actual world, or they sample it from their prior probability distribution. That is, each time people simulate a new counterfactual world, with probability s they make a copy of the actual world (they set the value of each variable to its actual-world value), and with probability $1 - s$ they sample the value of each exogenous variable from its probability distribution.

The second model is inspired by the *minimality* assumption, a common hypothesis in the literature on counterfactuals, which holds that people only consider counterfactuals that are 'minimally divergent' from what actually happened (e.g. Hiddleston, 2005; Rips, 2010). Therefore, the model assumes that people sample counterfactuals from their prior probability distribution, but keep only some of them, discarding those that are 'too far' from the actual world.

Specifically, we assume that people sample counterfactuals from their prior probability distribution, and divide these counterfactuals in two categories: those possible worlds in which the outcome of the game is similar to the actual world, and those in which

the outcome of the game is different from the actual world (i.e. worlds in which the player won the game, and worlds in which the player lost). Within each category, there is a subset of possible worlds that are the most similar to the actual world (here we define the similarity between two worlds as the number of variables that take the same value in both worlds). We assume that within each category, people only keep the possible worlds that belong to that subset, and that they discard the rest.

For example, in Study 2a, the player needs to draw at least two colored balls to win the game, and he draws a colored ball from all three urns. We assume that people simulate counterfactual possibilities by simulating possible rounds of the game from their probability distribution, and keep only the simulated worlds which are identical to the actual world, and those where the player draws a colored ball from only one urn (among the possible worlds where the player loses the game, these are the closest to what actually happened). Possible worlds where the player draws no colored balls at all are discarded because they are farther away from the actual world than worlds where the player draws one colored ball.

Neither model was able to reproduce people's causal judgments in studies 2, 3, or 4; see figures S10-12 in the Supplementary Information (https://osf.io/vnh84).

Note that we are not arguing that the XSM is necessarily the only counterfactual sampling model that could account for our experimental results, or even that the two models explored above are particularly reasonable alternatives. In exploring these alternative models, we simply highlight that our results cannot be explained by simply positing a general bias to sample counterfactuals that are both likely and close to the actual world. Our results also depend on the hypothesis one makes about the specific way that people make tradeoffs between probability and closeness.

General Discussion

Among the multitude of factors that contribute to an event, people spontaneously highlight one or a few of them as the event's cause(s). This suggests that people view some

factors as more causally responsible than others, and highlight these. In this paper, we have presented a theory of how people make these judgments of causal responsibility. Specifically, we argue that when judging to what extent C was the cause of E, people do the following:

a) They simulate counterfactual possibilities, i.e., alternative ways that the situation could have unfolded. They tend to simulate possibilities that are both likely and close to what actually happened (Lucas & Kemp, 2015).

b) They compute the causal strength of C for E by computing a statistical measure of effect size, such as the correlation between C and E across these counterfactuals (Quillien, 2020).

Our theory predicts people's causal judgments by combining two formal models. The Extended Structural Model of counterfactual reasoning (XSM; Lucas and Kemp, 2015) provides a model of how the human mind does (a). The Counterfactual Effect Size Model (CESM; Quillien, 2020) provides a model of how the mind does (b). In a reanalysis of existing data, and four new studies, we have shown that these models jointly provide a good account of people's causal judgments.

In the remainder of this paper, we discuss implications and limitations of this work, alternative theories of causal judgment, and directions for future research.

Causation and counterfactuals

At a broad level, our work is based on the idea that counterfactual reasoning is essential to intuitive judgments of causation. Our findings are therefore relevant to debates between counterfactual and non-couterfactual theories of causation.

Some philosophers reject counterfactual accounts of causation, favoring 'production' accounts according to which causation is a matter of transfer of force from the cause to the effect (Dowe, 1992; Salmon, 1994). Similarly, in psychology and linguistics the force dynamics framework holds that people make judgments of causation by consulting a

representation of the situation in terms of force vectors, akin to the forces in Newtonian mechanics (Talmy, 1988; Wolff, 2007). These theories have had some empirical success in explaining judgments of physical causation (Wolff, 2007). However, they do not account for other characteristics of human causal judgment (see e.g. Gerstenberg et al., 2021), and the current results illustrate these difficulties.

First we replicate and extend the finding that the prior probability of events influence causal judgments (Icard et al., 2017; Kominsky et al., 2015; Morris et al., 2019; Gerstenberg & Icard, 2020), a phenomenon that is difficult to explain on production accounts, but follows naturally from ours. Second, we find that we can influence people's causal judgments by manipulating the state of objects that are never in direct physical interaction with other objects during the event (as in Gerstenberg et al., 2021). In study 4, one of the flippers is not reached by the ball, but the orientation of that flipper has a large effect on people's judgments about whether the top flipper's orientation caused the player to win. We even find that the *probability distribution* over orientations of the unreached object influences participants' judgments.

Our data also speak to another alternative to counterfactual theories. Many existing findings supporting the counterfactual framework might be more parsimoniously explained by assuming that people engage in *hypothetical* instead of counterfactual reasoning. That is, people might make causal judgments about an event by consulting pre-existing simulations of what would happen, that they generated before the event actually happened (see Gerstenberg, 2022). Our new experimental results show that people simulate possibilities that are centered on what actually happened—effectively ruling out the hypothetical simulation hypothesis (see also Gerstenberg, 2022 for convergent evidence for causal judgments involving only one candidate cause).

Our theory builds on a large body of work exploring the connection between causal judgment and counterfactual reasoning (e.g. Morris et al., 2018; Gerstenberg et al., 2021; Lagnado et al., 2013; Wells & Gavanski, 1989; Halpern & Hitchcock, 2015; Khemlani et al.,

2014; Spellman, 1997; Petrocelli et al., 2011).[18] In particular, the idea that some counterfactual alternatives are more likely to be simulated than others plays a key role in Kahneman and Miller (1986) and many later accounts of causal judgment (Hitchcock & Knobe, 2009; Phillips et al., 2015).

The relationship between counterfactual sampling propensity and causal judgment is not trivial – for example, the events to which people most easily generate a counterfactual alternative are not always judged to be the most causally important (Mandel, 2003). As such, several formal theories have been designed to identify the function that maps counterfactual simulations to causal judgments (e.g. Morris et al., 2018; Halpern & Hitchcock, 2015; Spellman, 1997; Petrocelli et al., 2011). Our account is one such theory. In the present paper we have designed our experiments so that they could arbitrate between our model and the Necessity-Sufficiency Model (Icard et al., 2017), because these models stand out in their ability to explain recent empirical findings. In particular, other theories have not been able to accurately predict the particular ways in which the probabilities of events affect causal judgments. For example, many of these theories would predict that, everything else being equal, people will always favor unexpected events over expected events, or that people's causal intuitions are not sensitive to probabilistic considerations. But in many situations, people actually select expected events over unexpected events as causes (Kirfel et al., 2021; Quillien and German, 2021; Icard et al., 2017; O'Neill et al., 2021; Gerstenberg and Icard, 2020, see also Henne, Kulesza, et al., 2021; Henne et al., 2019). To our knowledge, only the CESM and NSM can explain this result.

In the next section we discuss one conceptual difference between our account and some prominent counterfactual theories of causal judgment.

---

[18] Note that some counterfactual theories of causal judgment have a slightly different scope than ours. For example, Gerstenberg et al. (2021) model how people make causal judgments when it is unclear that the outcome would have happened in the absence of the candidate cause; a key component of their theory is a model of how people estimate this uncertainty.

Necessity, Sufficiency, Pivotality, Criticality

Influential accounts of causation or responsibility judgment hold that these processes engage two qualitatively different kinds of counterfactual simulation: *retrospective* and *prospective* simulations (Icard et al., 2017; Gerstenberg et al., 2021; Lagnado et al., 2013). Retrospective simulation generates a counterfactual possibility by starting from what actually happened, and selectively modifying a few aspects of that particular situation. By contrast, prospective simulation generates a counterfactual possibility in a way that completely disregards the particulars of what actually happened. That is, when people engage in prospective simulation they simply sample a possibility from the prior probability distribution over ways that the situation could have unfolded.

For instance, one account of causal judgment holds that people separately compute whether an event was necessary for the outcome in these particular circumstances, and whether it is in general sufficient for the outcome (Icard et al., 2017). A theory of responsibility judgments assumes that people judge an agent to be responsible for an outcome if that agent was pivotal in this particular situation, and is in general critical for the outcome (Lagnado et al., 2013; Zultan et al., 2012; Chockler & Halpern, 2004).

These theories are motivated by an important observation: our sense of causation is sensitive to both the specifics of what actually happened and to general facts about the causal structure of the situation (Woodward, 2006; Lagnado et al., 2013; Woodward, 2021). For example, the lightning bolt was necessary for the forest fire in this one particular situation, and this contributes to people judging that it caused the fire. But people also care about whether a given factor reliably leads to the outcome in general—this is why they deny that the oxygen caused the fire, even though the fire would not have started in the absence of oxygen (Woodward, 2006; Lombrozo, 2010; Hitchcock, 2012; Icard et al., 2017; Kominsky et al., 2015; Vasilyeva et al., 2018).

Our account suggests an alternative explanation for why causal judgment is sensitive to both the specifics of what actually happened and general facts about causal

structure. We suggest that this pattern simply falls out of a very general fact about how people simulate counterfactual possibilities: people tend to simulate possibilities that are both likely and similar to what actually happened.

Good causes tend to be events that were necessary for the outcome in this particular situation, because counterfactual sampling is biased toward possibilities that are close to what actually happened. If C was necessary for E in the actual world, then in possible worlds that are close to the actual world, C and E will be highly correlated—because (for example) a counterfactual that is identical to the actual world except that C = 0 is also one where E = 0.

Good causes tend to be events that would reliably lead to the outcome across different possible situations, because counterfactual sampling is also biased toward a priori likely events. Because people simulate counterfactuals as a function of their prior probability, they tend to imagine a representative range of the different ways the situation could have unfolded. The correlation between C and E across these possible situations will tend to be high if C is sufficient for E across plausible background conditions.

In sum, under our account causal selection does not involve two qualitatively different kinds of counterfactual reasoning. Although causal selection tends to pick out causes that are necessary and sufficient (or pivotal and critical) for the outcome, this is a byproduct of a more general fact about how people sample counterfactuals; the notions of necessity, sufficiency, pivotality and criticality are not representational primitives that the mind computes explicitly.

Empirical data appear to favor the present account over theories that postulate two different kinds of counterfactual reasoning. In a re-analysis of studies of blame and responsibility attributions (Lagnado et al., 2013; Zultan et al., 2012), our theory predicts participants' judgments better than a model based on pivotality and criticality (Lagnado et al., 2013). We also tested the predictions made by the Necessity-Sufficiency model (NSM, Icard et al., 2017) in our experiments. We find that the NSM is able to reproduce

some key features of our data (especially in Study 3, where it is the best-fitting model), but that it fails to account for other important features[19].

Accounting for other documented features of human causal judgment

In this paper, we have focused of experimental paradigms which explicitly manipulate the prior probability of events, as well as the general causal structure of the situation and the specifics of what happened – these paradigms allow for the most direct tests of our account. Below we discuss how our account might explain a range of other features of human causal judgment[20].

Mental states

People are more likely to judge that an action caused an outcome when the agent intended that outcome (Lombrozo, 2010; Lagnado & Channon, 2008), and knew that the action would lead to the outcome (Kirfel & Lagnado, 2021; Lagnado & Channon, 2008).

These effects have a natural explanation under our framework. Suppose a boy breaks a vase by kicking a ball. Across different alternative ways that the event could have unfolded, is there a high correlation between the boy kicking the ball and the vase breaking? The answer depends on whether the boy intended to break the vase. If yes, we

[19] Note that we investigated only one specific version of the Necessity-Sufficiency hypothesis. The NSM is itself a flexible framework: in their original publication, Icard and colleagues (2017) provide a model for how people combine necessity and sufficiency when making causal judgments, but they do not commit to a specific formula for how the mind may compute necessity or sufficiency strength. To generate predictions from their model, we used the formulas suggested in their paper, but these formulas were only presented as suggestions. Future research may explore whether the NSM could accommodate the present data with suitably modified definitions of necessity and sufficiency strength. Alternatively, maybe necessity and sufficiency are computational primitives for causal judgment, but people combine them in a different way than suggested by the NSM.

[20] Note that many of these effects are compatible with other counterfactual theories. In particular, some of the results described below were first inspired by the Necessity-Sufficiency model.

expect that in counterfactual alternatives where the vase is slightly more to the left, the boy would have adjusted his aim accordingly, and still have broken the object (Lombrozo, 2010). By contrast, if the boy accidentally broke the vase, the link between him kicking the ball and the vase breaking was highly dependent on the idiosyncrasies of the particular circumstances: across the counterfactual possibilities we tend to consider, there is a low correlation between the boy kicking the ball and the vase breaking.

Similarly, suppose that the boy did not *know* that his kicking the ball would result in a broken vase. It is easy to imagine that, had he known, he would have acted differently (Kirfel & Lagnado, 2022). As such, the actions of ignorant agents are more weakly correlated with their outcomes across counterfactuals.

Double prevention

A bottle is about to fall. Peter is about to catch it, but Danielle accidentally knocks against him, making him unable to catch the bottle, which falls and breaks down. This is a case of *double prevention*, where D prevents P from preventing an effect E. In such cases, people are somewhat reluctant to say that D caused E (e.g. that Danielle caused the bottle to fall). This is sometimes considered a problem for counterfactual theories of causation, since E would not have happened in the absence of D. The CESM, however, gives a natural account of this intuition (O'Neill, Quillien, et al., 2022, see also Henne and O'Neill, 2022). Across possible counterfactuals to the event, the correlation between D and E is in general relatively low (especially in cases when D is not intentionally trying to bring about E), so the CESM predicts it should be assigned relatively low causal responsibility. In support of this explanation, judgments that D caused E increase when D is intentionally trying to bring about E (Lombrozo, 2010) and when participants are encouraged to think about counterfactual alternatives to D (O'Neill, Quillien, et al., 2022; Henne & O'Neill, 2022).

Robustness

Causal relationships are said to be *robust* (or *stable*, or *insensitive*) when they do not depend on the presence of moderating variables. For example, a drug that relieves headaches, but only in patients with a certain version of a gene, is not a robust cause of headache relief (Woodward, 2006). People generally judge that robust causal relationships are better causal relationships (Vasilyeva et al., 2018; Grinfeld et al., 2020; Nagel and Stephan, 2016; see also Phillips and Shaw, 2015; Murray and Lombrozo, 2017).

In general, robust causes tend to raise the probability of their effect more than non-robust causes. For example, a drug that relieves headaches no matter the patient's genotype will tend to have a higher overall probability of effectiveness than a drug that only works in patients with a particular version of a gene. In such cases, robust causes are more highly correlated with their effect, across counterfactuals, than non-robust ones, and so our account naturally predicts a preference for the former. More generally, the CESM predicts that causes are better if they tend to lead to the effect across most possible background circumstances. Indeed, if A causes E but only when B is present, people agree that A caused E to the extent that B was a priori likely (Kominsky et al., 2015; Morris et al., 2019).

However, robustness can also be manipulated independently of probability-raising; this happens if for example the drug in the 'non-robust cause' condition is extremely effective in patients that have the right version of the gene, and the drug in the 'non-robust cause' condition is mildly effective across the board. Such cases are slightly more complicated both theoretically and empirically – we briefly discuss them in the Supplementary Information.

Morality, recency, and omission effects

Consider the following three findings about human causal judgment:

- People tend to view moral norm violations as more causal than their

norm-conforming counterparts; intuitively we think that the car crash at the intersection was caused by the driver who crossed the red light and not the one who crossed at the green light (Knobe & Fraser, 2008; Hitchcock & Knobe, 2009; Roxborough & Cumby, 2009; Alicke, 1992; Willemsen & Kirfel, 2019).

- People tend to think of actions as more causal than omissions. For example, it feels natural to say that a gardener who uprooted a plant caused the plant to die, but we are less likely to say so if the gardener simply failed to water the plant (Walsh & Sloman, 2011; Cushman & Young, 2011).

- People privilege recent events over earlier ones. Suppose a basketball player scores a shot that allows his team to take the lead at the very last second of the game. Intuitively, the last-second shot caused the team to win the game, and is more responsible for the victory than points scored earlier in the game (Ziano & Pandelaere, 2022; Cusick & Peter, 2015).

Counterfactual accounts suggest that these three classes of effect have fundamentally the same explanation. The explanation relies on the fact that many factors, besides the ones we modeled here (prior probability and actual-world closeness), have been found to influence counterfactual reasoning. In particular, there is independent empirical evidence (reviewed in Byrne, 2016) that when people simulate counterfactuals, they are more likely to:

- replace a norm violation with a norm-conforming action than vice-versa,

- generate alternatives to recent compared to early events,

- replace an action with an omission than vice-versa.

Therefore, a natural hypothesis is that morality, time and action / omission asymmetries affect causal judgment via their effect on counterfactual simulation. This

hypothesis can be tested by probing people's intuitions in *disjunctive* cases, where each of two causes would have been individually sufficient for the outcome. Counterfactual accounts predict that in such cases, the effects described above should *reverse*.[21] That is, in a case where each one of two events would have been individually sufficient to lead to the outcome, people will think that:

- norm-conforming actions are more causal than norm-violating actions,
- early events are more causal than late events,
- omissions are more causal than actions.

Each of these reversals does in fact happen (see Icard et al., 2017 for injunctive norms; Henne et al., 2019 for omissions; Henne, Kulesza, et al., 2021 for recency). For example, suppose that a computer has a bug such that if anyone logs in, some emails will be deleted; and two employees happen to log in at the exact same time. When judging who caused the email deletion, people view the employee who had permission to log in as more causal than the employee who did not have permission (Icard et al., 2017). Non-counterfactual accounts of the effect of moral norm violations on causal judgment[22] fail to predict this effect.

[21] On our account, this is for the same reason that the effect of prior probability reverses in disjunctive compared to conjunctive structures, see Figures 5 and 6 earlier. In disjunctive structures, events with high counterfactual sampling propensities are more highly correlated to the outcome, and therefore are viewed as more causal. Assuming that norm-conforming events have higher sampling propensity, our account naturally predicts that they will be viewed as more causal in a disjunctive case. See also Icard et al. (2017) for an alternative account of the reversal.

[22] On these accounts, our motivation to blame might distort our causal judgments (Alicke et al., 2011), our intuitive concept of causation might be inherently normative (Sytsma, 2021), or there might be pragmatics confounds in the experimental tasks that researchers use (Samland & Waldmann, 2016). For other experimental findings that cast doubt on non-counterfactual explanations of the effect of moral violations on causal judgment, see Hitchcock and Knobe, 2009; Kominsky and Phillips, 2019; Phillips et al., 2015.

Judgments of blame and responsibility

Our account views causation as a relationship between *events*, but people also routinely assign causal responsibility to *agents*. Our model is able to account for some of these judgments: we were able to accurately predict people's judgments about whether a player was to blame, or responsible, for their team's defeat (Lagnado et al., 2013; Zultan et al., 2012), by asking our model whether the team lost because of the fact that the player failed. This result suggests that the intuitive conception of causation we study here is a core component of the way we attribute blame and responsibility to agents (see also Phillips and Shaw, 2015; Cushman and Young, 2011; Pizarro et al., 2003; Cushman, 2008). But we think this is only one piece of the puzzle; the full information-processing logic of blame and responsibility judgments is still a topic of active research (e.g. Langenhoff et al., 2021).

Levels of analysis

Our account gives a functional characterization of human causal judgment, but does not entail strong commitments about process-level details. We remain agnostic about the mechanisms via which the mind samples counterfactuals, and how the process is implemented neurally. A growing literature has explored ways in which human cognition is resource-rational: in many tasks, people use smart sampling strategies which allow them to make good enough estimates by taking only a small number of samples (Vul et al., 2014; Lieder et al., 2018). An open challenge for future research is to discover how people make causal judgments efficiently by sampling a realistic number of counterfactual possibilities (see Bramley et al., 2017; Davis and Rehder, 2020 for sampling-based process models of other domains of causal cognition).

Notably, counterfactual simulation is only tractable if people consider a small subset of the potentially relevant variables. For instance, when we judge what caused the plants in our apartment to die, we probably do not simulate the possibility that the Pope could have watered them, even though the plants would have survived if he had. This raises the

question of how people decide which variables to simulate in the first place – see Morris et al. (2021) for a similar issue in decision-making.

At a higher level of abstraction, cognitive scientists often aspire to explain cognitive processes by specifying the information-processing problem that they solve, and deriving good solutions to that problem (Marr, 1982; Anderson, 1990; Cosmides & Tooby, 1994). To a large extent, the current work is inspired by that approach. We follow researchers who hold that causal judgment is well-designed to identify robust causal relationships (Lombrozo, 2010; Hitchcock, 2012; Woodward, 2021) and good points of intervention (Morris et al., 2018). Effect size computations that range across counterfactuals are a sensible solution to that problem. Standardized effect size statistics are widely used in science, and have the advantage of being invariant to the particular unit of measure being used (e.g. the effect of temperature does not depend on whether it was measured in Fahrenheit or Celsius). It is also plausible that causal judgments are designed to convey information about the specifics of what happened, in addition to information about the general strength of a causal relationship. This implies that the mind needs to somehow integrate both types of information. The model of counterfactual sampling that we used here – which specifies one way such an integration can be done – can be derived from normative considerations (see Appendix in Lucas and Kemp, 2015)[23]. At the same time, a full specification of the information-processing problem the mind is solving when making causal judgments (and of an optimal solution to that problem) remains an open problem for future research.

[23] For example, if we view counterfactual simulation as the process of re-winding the world back to a certain point back in time, and then playing it back, we might assume that there is a small probability p at each instant for each variable to take a different path than the one it took in the actual world. This assumption can be shown to lead to the form of the re-sampling process used in the XSM.

Limitations and directions for future research

Here we analyzed existing studies, and conducted new experiments, that elicited causal judgments for diverse kinds of causal structures. Our account reproduces people's intuitions in classical disjunctive and conjunctive structures with 2 causal variables (Morris et al., 2019; O'Neill et al., 2021), but also in more complicated settings (Quillien and Barlev, 2022; Lagnado et al., 2013; Zultan et al., 2012, and our new experiments).

Future research should extend this investigation to other kinds of causal structures. In particular, while we focused on situations where all causal links are direct, people often have to make judgments about causal chains (of the form $A \rightarrow B \rightarrow C$, see Lagnado and Channon, 2008; Nagel and Stephan, 2016; Johnson and Ahn, 2015). With the exception of Quillien and Barlev (2022), we also looked exclusively at causal structures where there is no potential confounding between variables (i.e. variables upstream of the outcome variable are statistically independent from each other). In such causal structures, the CESM is particularly simple, because it predicts that the causal strength of C for E is simply the correlation between C and E across counterfactuals. Nonetheless, the model also makes predictions in the more general case, and they could fruitfully be tested.

The model of counterfactual sampling we use here (the Extended Structural Model, Lucas and Kemp, 2015) is not a comprehensive psychological model of counterfactual sampling. The model specifies how the prior probability of an event, as well its similarity to what actually happened, influence which counterfactual alternatives people simulate; but many other factors (such as moral norms and epistemic states) influence counterfactual simulation (Kahneman & Miller, 1986; Kirfel & Lagnado, 2022; Byrne, 2016). To fully understand causal selection, we would need to know how people integrate all these different factors when they decide which counterfactual possibilities to consider.

The XSM has previously been empirically tested only in settings where events can easily be modeled as binary variables. This was not a problem for our purposes, since our experiments only concerned systems that can be modeled in this way. However, many

causal judgments seem to involve continuous variables, as in the example "the harvest was good this year because of the high temperatures". Future research could investigate how people think about counterfactuals in situations involving continuous variables, and explore implications for formal models of counterfactual and causal reasoning.

Conclusion

An increasingly popular idea in cognitive science is that when people judge whether an event caused another, they think about alternative ways things could have happened. But for everything that happens, there are an infinity of possible ways that the situation could have unfolded differently than it did. Which of these possibilities come to people's minds when they make causal judgments?

Here we suggested that the answer might lie in existing work on how people reason about counterfactual possibilities in general. When people imagine alternatives to something that happened, they tend to think of possibilities that are relatively similar to what actually happened, and a priori likely.

Our experimental results suggest that people sample counterfactual possibilities in the same way when they make causal judgments. We also find support for the idea that people make causal judgments by computing a statistical measure of effect size—such as the correlation between the candidate cause and the outcome—across these counterfactual possibilities.

Some cognitive scientists have proposed that to explain human causal judgment, we need to posit that people reason about abstract features of the relationship between the cause and the effect, such as whether the candidate cause was necessary and sufficient for the effect. The current results suggest an alternative account: people's causal intuitions emerge naturally from simple facts about the counterfactual possibilities that come to mind, and the way we compute the dependence of the outcome on the cause across these possibilities.

References

Ackerman, N. L., Freer, C. E., & Roy, D. M. (2011). Noncomputable conditional distributions. *2011 IEEE 26th Annual Symposium on Logic in Computer Science*, 107–116.

Ahn, W.-k., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*(3), 299–352.

Alicke, M. D. (1992). Culpable causation. *Journal of personality and social psychology*, *63*(3), 368.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670–696.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

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., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

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

Byrne, R. M., & Johnson-Laird, P. N. (2020). If and or: Real and counterfactual possibilities in their truth and probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(4), 760.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, *37*(6), 1171–1191.

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

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

Cosmides, L., & Tooby, J. (1994). Beyond intuition and instinct blindness: Toward an evolutionarily rigorous cognitive science. *Cognition*, *50*(1-3), 41–77.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075.

Cusick, C., & Peter, M. (2015). The last straw fallacy: Another causal fallacy and its harmful effects. *Argumentation*, *29*(4), 457–474.

Danks, D. (2017). Singular causation. *The oxford handbook of causal reasoning*, 201–215.

Davis, Z. J., & Rehder, B. (2020). A process model of causal reasoning. *Cognitive Science*, *44*(5), e12839.

De Brigard, F., Henne, P., & Stanley, M. L. (2021). Perceived similarity of imagined possible worlds affects judgments of counterfactual plausibility. *Cognition*, *209*, 104574.

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

Demirtas, H. (2022). Causation comes in degrees. *Synthese*.

Dowe, P. (1992). Wesley salmon's process theory of causality and the conserved quantity theory. *Philosophy of science*, *59*(2), 195–216.

Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals, and causal judgments. *Philosophical Transactions of the Royal Society B*.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35).

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological review*, *128*(5), 936.

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

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

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford handbook of causal reasoning*, 515–548.

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

Gill, M., Kominsky, J. F., Icard, T. F., & Knobe, J. (2022). An interaction effect of norm violations on causal judgment. *Cognition*, *228*, 105183.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2014). *Concepts in a probabilistic language of thought* (tech. rep.). Center for Brains, Minds and Machines (CBMM).

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological review*, *111*(1), 3.

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.

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

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

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 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*.

Hanley, B. J. (2021). What caused the bhopal gas tragedy? the philosophical importance of causal and pragmatic details. *Philosophy of Science*, *88*(4), 616–637.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law*. OUP Oxford.

Henne, P., Kulesza, A., Perez, K., & Houcek, A. (2021). Counterfactual thinking and recency effects in causal judgment. *Cognition*, *212*, 104708.

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

Henne, P., & O'Neill, K. (2022). Double prevention, causal judgments, and counterfactuals. *Cognitive Science*, *46*(5), e13127.

Henne, P., O'Neill, K., Bello, P., Khemlani, S., & De Brigard, F. (2021). Norms affect prospective causal judgments. *Cognitive Science*, *45*(1), e12931.

Hesslow, G. (1988). The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality* (pp. 11–32).

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

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

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

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

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587–612.

Icard, T. F. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

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

Johnson, S. G., & Ahn, W.-k. (2015). Causal networks or causal islands? the representation of mechanisms and the transitivity of causal judgment. *Cognitive science*, *39*(7), 1468–1503.

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

Kaiserman, A. (2016). Causal contribution. *Proceedings of the Aristotelian Society*, *116*(3), 387–394.

Kaiserman, A. (2018). 'more of a cause': Recent work on degrees of causation and responsibility. *Philosophy Compass*, *13*(7), e12498.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist*, *28*(2), 107.

Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in human neuroscience*, *8*, 849.

Kirfel, L., Icard, T., & Gerstenberg, T. (2021). Inference from explanation. *Journal of Experimental Psychology: General*.

Kirfel, L., & Lagnado, D. (2021). Causal judgments about atypical actions are influenced by agents' epistemic states. *Cognition*, *212*, 104721.

Kirfel, L., & Lagnado, D. (2022). Changing minds—epistemic interventions in causal reasoning.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. *Moral psychology*, *2*, 441–447.

Knobe, J., & Shapiro, S. (2021). Proximate cause explained. *The University of Chicago Law Review*, *88*(1), 165–236.

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

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

Krasich, K., O'Neill, K., & De Brigard, F. (n.d.). Looking at mental images: Eye-tracking mental simulation during retrospective causal judgment.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770.

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

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

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

Lewis, D. (1973a). Causation. *The journal of philosophy*, *70*(17), 556–567.

Lewis, D. (1973b). *Counterfactuals*. John Wiley & Sons.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

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

Lucas, C., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Lucas, C., & Kemp, C. (2012). A unified theory of counterfactual reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34).

Lucas, C., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological Review*, *122*(4), 700.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Mill, J. S. (1843). *A system of logic ratiocinative and inductive: 1* (Vol. 2). Longmans.

Morris, A., Phillips, J., Gerstenberg, T., & Cushman, F. (2019). Quantitative causal selection patterns in token causation. *PloS one*, *14*(8), e0219704.

Morris, A., Phillips, J., Huang, K., & Cushman, F. (2021). Generating options and choosing between them depend on distinct forms of value representation. *Psychological Science*, *32*(11), 1731–1746.

Morris, A., Phillips, J., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). Judgments of actual causation approximate the effectiveness of interventions.

Murray, D., & Lombrozo, T. (2017). Effects of manipulation on attributions of causation, free will, and moral responsibility. *Cognitive science*, *41*(2), 447–481.

Nagel, J., & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. *Proceedings of the 38th annual conference of the cognitive science society*, 806–812.

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

O'Neill, K., Henne, P., Bello, P., Pearson, J., & De Brigard, F. (2022). Confidence and gradation in causal judgment. *Cognition*, *223*, 105036.

O'Neill, K., Henne, P., Pearson, J., & De Brigard, F. (2021). Measuring and modeling confidence in human causal judgment. *Advances in neural information-processing systems*.

O'Neill, K., Quillien, T., & Henne, P. (2022). A counterfactual model of causal judgment in double prevention. *Conference in computational cognitive neuroscience*.

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

Pearl, J. (2000). *Causality*. Cambridge university press.

Pearl, J. (2013). Structural counterfactuals: A brief introduction. *Cognitive science*, *37*(6), 977–985.

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

Petrocelli, J. V., Percy, E. J., Sherman, S. J., & Tormala, Z. L. (2011). Counterfactual potency. *Journal of personality and social psychology*, *100*(1), 30.

Pfeifer, N., & Tulkki, L. (2017). Conditionals, counterfactuals, and rational reasoning: An experimental study on basic principles. *Minds and Machines*, *27*(1), 119–165.

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.

Phillips, J., & Shaw, A. (2015). Manipulating morality: Third-party intentions alter moral judgments by changing causal reasoning. *Cognitive Science*, *39*(6), 1320–1347.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of experimental social psychology*, *39*(6), 653–660.

Quillien, T., & Lucas, C. (2023). *Supplementary material for "counterfactuals and the logic of causal selection", open science framework: https://osf.io/h42f7/?view_only=697cc7504ee345a799615cef2d260c01* (tech. rep.).

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

Quillien, T., & Barlev, M. (2022). Causal judgment in the wild: Evidence from the 2020 us presidential election. *Cognitive Science*.

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*, 104806.

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

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

Roxborough, C., & Cumby, J. (2009). Folk psychological concepts: Causation. *Philosophical Psychology*, *22*(2), 205–213.

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

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

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

Sartorio, C. (2020). More of a cause? *Journal of Applied Philosophy*, *37*(3), 346–363.

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

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

Sloman, S., & Lagnado, D. (2015). Causality in thought. *Annual review of psychology*, *66*, 223–247.

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

Stalnaker, R. C. (1981). A theory of conditionals. *Ifs: Conditionals, belief, decision, chance and time* (pp. 41–55). Springer.

Stanley, M. L., Stewart, G. W., & Brigard, F. D. (2017). Counterfactual plausibility and comparative similarity. *Cognitive Science*, *41*, 1216–1228.

Starr, W. (2019). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019). Metaphysics Research Lab, Stanford University.

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation—a computational model. *Cognitive Science*, *44*(7), e12871.

Sytsma, J. (2021). Causation, responsibility, and typicality. *Review of Philosophy and Psychology*, *12*(4), 699–719.

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

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

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

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

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgements and norms. *Philosophy Compass*, *14*(1), e12562.

Wolff, P. (2007). Representing causation. *Journal of experimental psychology: General*, *136*(1), 82.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford university press.

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

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

Zhao, B., Lucas, C. G., & Bramley, N. R. (2021). How do people generalize causal relations over objects? a non-parametric bayesian account. *Computational Brain & Behavior*, 1–23.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological review*, *127*(5), 719.

Ziano, I., & Pandelaere, M. (2022). Late-action effect: Heightened counterfactual potency and perceived outcome reversibility make actions closer to a definitive outcome seem more causally impactful. *Journal of Experimental Social Psychology*, *100*, 104290.

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