Productive causation and compositionality

Tadeg Quillien[1], Neil Bramley[1], and Christopher G Lucas[2]

[1]Department of Psychology, University of Edinburgh

[2]School of Informatics, University of Edinburgh

Author Note

We have no conflicts of interest to report. Data and analysis code for all studies are available on the Open Science Framework at https://osf.io/nsfx6/?view_only=3c9816af19ab463f9e9390740d622d86. Studies 2 and 3 were pre-registered (Study 2: https://osf.io/a3prm/?view_only=208fc61e8e6644cf81eb513a95a69d06, Study 3: https://osf.io/cya8k/?view_only=6fc28c92f2b5405bb5652d6232ee2fdf).

Abstract

According to counterfactual theories of causal judgment, people judge causation by evaluating the consequences of counterfactual interventions on their representations of the world. What is the format of these representations? Research on causal generalization suggests that people hold ‘invariant’ representations of causal relationships that can be flexibly composed together. We suggest that people can compute counterfactuals on the basis of these invariant representations, and in particular they can imagine ‘disconnecting’ the causal link between two variables. Importantly, this hypothesis implies that causal judgment is supported by richer representations than the Structural Causal Models (SCMs) used in most counterfactual theories. We argue that this gap can explain the shortcomings of existing counterfactual theories, for example why these theories struggle to explain the distinction between ‘productive’ and non-productive causation. In a series of simple experiments, we find that the consequences of variable-disconnection counterfactuals systematically affect people’s causal judgments, even holding constant the structural causal model describing the situation. Overall, the counterfactual framework might provide a unifying account of human causal judgment, provided we correctly understand the mental representations people use to imagine counterfactual scenarios.

Keywords: causality; causal reasoning; generalization; compositionality

Productive causation and compositionality

Introduction

According to an influential family of accounts, people judge causation by engaging in counterfactual reasoning. Consider for example the following scenario:

Rock-throwing. Suzy and Billy are playing in the garden. Suzy throws a rock at a nearby bottle. Billy, standing in the path of the rock, could easily catch it but decides not to. Suzy's aim is perfectly accurate, and the bottle breaks.

It seems reasonable to say:

(1) The bottle broke because Suzy threw a rock at the bottle.

But it also seems fairly reasonable to say:

(2) The bottle broke because Billy did not catch the rock.

According to the counterfactual framework, we think (1) and (2) because we think that:

(1') If Suzy had not thrown a rock, the bottle would not have broken.

(2') If Billy had caught the rock, the bottle would not have broken.

Counterfactual theories have successfully explained a wide range of phenomena in causal cognition. This success is largely due to their use of a precise formalism for representing causal relationships: Structural Causal Models (SCMs, Pearl, 2000). A structural causal model represents a causal system in terms of variables and the structural relationships between them. Many counterfactual accounts of causal reasoning assume, implicitly or explicitly, that people represent a given causal system by constructing a causal model of that system, and that they derive causal judgments from that causal model.

We argue that this assumption is problematic: People probably represent causal relationships using richer representations than structural causal models. A causal model is designed to represent a particular causal system and typically discards some information about the more general causal laws that explain why the system behaves the way it does (Tenenbaum et al., 2007; Griffiths and Tenenbaum, 2009, see also Maudlin, 2004). We argue that causal judgments are (at least sometimes) derived from a representation of these more general causal laws.

This hypothesis sheds light on phenomena that counterfactual theories typically struggle to explain. It is difficult for example to resist the intuition of a deep qualitative difference between the respective causal roles of Suzy and Billy:

(1") Suzy's throw *caused* the bottle to break.

(2") Billy's inaction *allowed* the bottle to break.

Since changing either Suzy's throw or Billy's inaction would have prevented the bottle from breaking, it is challenging to derive a qualitative difference between their causal contributions from a counterfactual framework (Hall, 2004; Hitchcock, 2007). We argue that the intuition of a difference arises from a process of compositional causal reasoning, operating over the causal laws with which people represent the situation. To preview our argument, the difference between Suzy and Billy is related to the fact that only the first of these two counterfactual statements is true:

(1"') If Suzy had not been there, the bottle would not have broken.

(2"') #If Billy had not been there, the bottle would not have broken.

These counterfactuals, in which we mentally 'disconnect' some variables from the system, are difficult to represent using the formalism of structural causal models, but they can be evaluated using richer representations of the relevant causal laws. In a series of simple experiments, we show that manipulating the consequences of these 'variable-disconnection' counterfactuals has an efffect on people's causal judgments, even holding constant the causal models representing the situation.

Implications

If our ideas are on the right track, they ultimately support the counterfactual approach to causation, by offering an explanation for data that seem to create difficulties for the approach. Because counterfactual theories struggle to account for qualitative differences in causal intuitions (like the difference between Billy and Suzy), some researchers hold that we might have two different concepts of causation (Hall, 2004; Lombrozo, 2010). In addition to a *counterfactual* concept of causation, we might also have a *process*, or *productive* concept, defining causation in terms of whether there is a continuous physical process that goes all the way from the cause to the outcome (Dowe, 1992; Salmon, 1994; Wolff, 2007).

We argue that the data that motivate causal pluralism can be more parsimoniously explained within a purely counterfactual framework, provided we make correct assumptions about the representations from which people compute counterfactuals.

Scope of the work

In this paper we use the term 'causal judgment' to refer to judgments of singular causation, as opposed to other cognitive operations such as causal learning or inference. That is, we assume that reasoners already know how the relevant causal system works and what happened in the situation of interest, and must choose how to describe what caused an event. This problem itself consists of sub-problems such as binary judgments of actual causation (is event C a cause of outcome E or not?, e.g. Halpern and Pearl, 2005), quantitative attribution of causal responsibility (is C a more important cause than other causes of E?, e.g. Quillien and Lucas, 2023), and choice of causal verb (did event C cause E or merely allowed E to happen?, e.g. Wolff et al., 2010). Our arguments are potentially relevant to all three sub-problems; in our experiments we will mostly focus on the latter one, as causal verb selection is often used to study judgments of productive causation (e.g. Wolff, 2007; Beller & Gerstenberg, 2023).

The standard counterfactual approach

According to the counterfactual approach to causation, 'C caused E' means (roughly) that if C had not been the case, then E would not have been the case (e.g. Lewis, 1973).[1] Most modern implementations of this idea use the formalism of Structural Causal

[1] We will often use letters like C to denote variables. Abusing notation slightly (as is common in the literature) we will abbreviate $C = 1$ as C when context makes clear what we mean.

Models (SCMs)[2] to specify this idea in a computationally precise manner (for exceptions see e.g. Goldvarg and Johnson-Laird, 2001; Gerstenberg et al., 2021; Wolff et al., 2010). We collectively refer to counterfactual theories using this formalism as belonging to the 'standard counterfactual framework' (for examples see Hitchcock, 2001; Woodward, 2003; Menzies, 2004; Halpern and Pearl, 2005; Lagnado et al., 2013; Gallow, 2021; Quillien and Lucas, 2023).

Structural Causal Models

Structural Causal Models (SCMs) are formal objects that encode knowledge about causal relationships, and can be used for various purposes such as inference, counterfactual reasoning, and decision-making (Spirtes et al., 1993; Pearl, 2000). SCMs are a particular type of causal graphical models; they are closely related to Causal Bayes Nets, which also encode causal information but are less useful for computing counterfactuals (Pearl, 2000). Our introduction will be relatively informal; for extended treatment we refer the reader to Pearl (2000) or Halpern (2016).

An SCM is a representation of a causal system in terms of *variables*, and *structural equations* that determine how the value of a variable is determined by the value of other variables. Consider our rock-throwing scenario:

Rock-throwing. Suzy and Billy are playing in the garden. Suzy throws a rock at a nearby bottle. Billy is standing on the path of the rock, and could easily catch it, but he decides not to. The bottle breaks.

We can represent the causal structure of the situation in terms of *variables* and *structural equations.* Variable R represents whether Suzy throws her rock, C whether Billy catches the rock, and B represents whether the bottle breaks; for simplicity we treat all

[2] Structural Causal Models are sometimes also called Structural Equation Models or Functional Causal Models. Our arguments about the limitations of SCMs also apply to the closely related formalism of Causal Bayes Nets. We focus here on SCMs because Causal Bayes Nets are generally not considered a promising substrate for representing counterfactuals (Pearl, 2000).

variables as binary (e.g. R is 1 if Suzy throws the rock and 0 otherwise). R and C are *exogenous* variables, in the sense that we do not explicitly represent their causes. In contrast, B is *endogenous*, since its value depends on the value of R and C. We express this dependency with a structural equation:

$$B := R\&\neg C$$

This equation says that the bottle breaks if and only if Suzy throws the rock ($R = 1$) and Billy does not catch it ($C = 0$). The := operator differs from a standard equality operator because it encodes the asymmetry of causation: the equation cannot be re-arranged in a way that would make Suzy's throw causally dependent on whether the bottle breaks, for example. R and C are said to be the *parents* of B, since they appear in the structural equation for B.

Intervening on a variable consists in 'cutting' that variable from the influence of its parents and setting it to a value of our choice. For example we could intervene to force the bottle to remain intact, regardless of whether Suzy throws the rock, by replacing the structural equation for B with:

$$B := 0$$

Interventions are key to formalizing the asymmetry of causal relationships: since R is a parent of B, an intervention on R would have an effect on B, but not vice-versa.

Finally, we can represent the actual situation described in the scenario (where Suzy throws the rock, Billys doesn't catch it, and the bottle break) by assigning values to the variables: in the actual world, we have $R = 1$, $C = 0$, and it follows from the structural equation that $B = 1$.

SCMs, and causal graphical models more generally, have been usefully applied to many aspects of human causal cognition, such as learning, inference, decision-making and categorization (for reviews see e.g. Gopnik et al., 2004; Sloman & Hagmayer, 2006; Holyoak & Cheng, 2011; Rehder, 2017). This makes causal models an attractive candidate as the representational substrate of causal judgments.

Counterfactuals and structural causal models

Re-expressed in the language of SCMs, the counterfactual theory of causation can be expressed in schematic form as follows:

'Suppose that in the actual world, $C = c$ and $E = e$. If an intervention setting C to a different value than c would result in E taking a different value than e, then $C = c$ caused $E = e$.'

For example, intervening to prevent Suzy from throwing the rock (setting R to $R = 0$), prevents the bottle from breaking ($B = 0$). Therefore Suzy's throw ($R = 1$) is a cause of the bottle breaking ($B = 1$). Analogously, Billy's inaction is also a cause of the bottle breaking.

This basic insight has been elaborated upon in many ways by philosophers, computer scientists and psychologists. For example, researchers have used the formalism of SCMs to develop counterfactual theories that give the right verdict even in cases where there is no direct counterfactual dependence between the outcome and the cause, as when several soldiers simultaneously shoot a prisoner (Menzies, 2004; Halpern & Pearl, 2005; Hall, 2007; Hitchcock, 2001; Gallow, 2021; Beckers & Vennekens, 2018). They have also developed theories that account for gradation in judgments of causal responsibility (Chockler & Halpern, 2004; Lagnado et al., 2013; Icard et al., 2017; Quillien, 2020; Quillien & Lucas, 2023). These theories have a good track record of empirical success (e.g. Kominsky et al., 2015; Gerstenberg & Icard, 2020; Henne et al., 2021; Quillien & Barlev, 2022; Gill et al., 2022; O'Neill et al., 2022; Konuk et al., 2023; Xiang et al., 2023).

Empirical limitations of the standard counterfactual framework.

Remember that there seems to be a qualitative difference between the causal role of Suzy and that of Billy in our rock-throwing scenario. Empirically, the fact that the mind makes such a distinction is suggested by the way people talk about causation (Pinker, 2007). For example, we can use causative verbs to describe Suzy's causal role, but not Billy's (Rose et al., 2021):

(1”) Suzy’s throw broke the bottle

(2”) #Billy’s inaction broke the bottle

(# signs indicate statements that sound a bit off). There is also a double dissociation between the acceptability of the verbs ‘Cause’ and ‘Allow’ (Wolff, 2007; Wolff et al., 2010; Livengood & Machery, 2007; Walsh & Sloman, 2011; Thanawala & Erb, 2024):

(1”’) Suzy’s throw caused the bottle to break.

(2”’) #Billy’s inaction caused the bottle to break.

(1””) #Suzy’s throw allowed the bottle to break.

(2””) Billy’s inaction allowed the bottle to break.

In sum, while Billy’s inaction is a cause of the bottle breaking, it seems like it is a different, weaker kind of cause than Suzy’s throw. The existence of these sorts of intuitions motivates the thesis of causal pluralism, according to which our mind hosts two different concepts of cause (Hall, 2004; Lombrozo, 2010). Causal pluralism holds that the mind has a notion of *productive causation*, instantiated when a cause produces an outcome by transmitting a physical quantity (as in when Suzy breaks the bottle by throwing the rock), and a weaker notion of *dependence causation*, where an outcome is counterfactually dependent on the cause but there is no physical connection between them (as when Billy allows the bottle to break by failing to catch the rock).

It seems difficult to account for the distinction between productive and non-productive causation in terms of counterfactual interventions over causal models. A counterfactual account seems bound to treat Billy and Suzy’s behavior in a symmetrical manner, because the bottle would have remained intact if we had performed an intervention on either variable (Hitchcock, 2007).

One possible way to rescue a counterfactual account would be to argue that Billy’s failure to catch the rock is not really an event but an ‘omission’, and that omissions cannot count as true causes (Beebee, 2004). However, an intuitive distinction between productive and non-productive causation also arises between two non-omissive causes, as illustrated by a case of ‘double-prevention’:

Double prevention: Suzy and Billy are playing in the garden. Suzy throws a rock at a nearby bottle. Billy is about to catch the rock, which would prevent it from breaking the bottle. However, Danielle pushes Billy away, preventing him from preventing the bottle from breaking. The bottle breaks.

Danielle’s pushing Billy away is a concrete event rather than an omission, yet people typically view her action as a non-productive cause of the outcome (McGrath, 2003; Hall, 2004; Lombrozo, 2010; Rose et al., 2021; Thanawala & Erb, 2024).

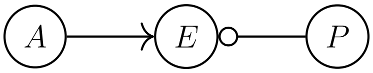
SCM-based accounts also face difficulties when modeling judgments of *actual causation*, i.e. modeling how people make a binary distinction between causes and non-causes of an outcome (Halpern, 2016). For example Hall (2007) constructed pairs of scenarios that are *structural isomorphs*: these scenarios can be represented by the same SCM, but in one scenario we clearly have the intuition that the target event caused the outcome, while in the other scenario we have the intuition that it did not (we give an example in the Supplementary Information). Most theorists think these cases are concerning; when they attempt to give formal accounts of actual causation they typically use additional representational machinery beyond bare SCMs—like a distinction between ‘default’ and ‘deviant’ states of a variable (Menzies, 2004; Hall, 2007; Hitchcock, 2007; Gallow, 2021, but see Blanchard and Schaffer, 2017; Wysocki, 2023).

In the next section, we explore other limitations of structural causal models. At first sight, the material that we review there is motivated by quite different considerations than the issues we just discussed. Instead of talking about how people make judgments of singular causation, we will discuss what makes people good at generalizing causal information. We will eventually circle back to where we started, arguing that this work is relevant to questions about causal judgment.

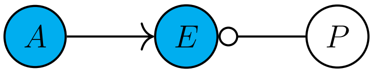
Compositionality and causal representation

Machines

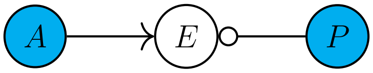
In this section we will illustrate our arguments with examples from a toy universe of 'machines' with simple causal rules.[3] Consider the following system:



This diagram represents a simple device where 'nodes' are connected to each other. Nodes can be active (blue) or inactive (white). A node is activated if it receives an input from at least one active node connected with a 'generative' link (represented as ⟶) and does not receive any input from an active node connected with a 'preventative' link (represented as —○). A node that does not receive any input can also be activated by an external intervention. So here node E fires if it receives a stimulation from A and is not inhibited by P. If we turn A on, it will activate E:



But if we turn P on, this de-activates E:

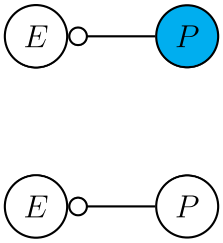


In these diagrams, the arrows (⟶ and —○) contain more information than in the simple Directed Acyclic Graphs (DAG) that are often used to represent causal models. In a DAG, an arrow between two variables means that the two variables are causally related, but it does not specify the nature of this causal influence. The ⟶ and —○

[3] These systems are usually called 'neuron diagrams' by philosophers. We instead call these systems 'machines' to avoid possible ambiguity among our psychologist readers.

links we use here contain substantial information about the causal relationship, specifying whether it is generative or preventative. On the other hand, it is important to note that the rules of our toy universe of machines are meant as illustrative examples: they do not embody psychological claims. That is, we are not making any claims about how the human mind represents 'generative' and 'preventative' causation in general; these terms are here defined purely in terms of their functional role within the machine universe.

To preview our argument, consider a machine where node E only receives input from a preventative node P. Then E is always Off regardless of the state of P, since there is nothing that could activate E:



So a Structural Causal Model describing this machine is simply:

$$E := 0$$

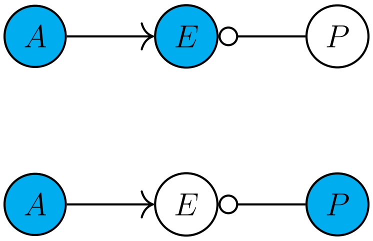
Intuitively this SCM leaves out important information about the machine, so there might be something wrong with SCMs as a substrate for causal representation.

Invariant causal representation

Human causal cognition is highly compositional: People can understand how a system works in terms of how its sub-parts are combined, along with the causal laws governing the relevant domain (Cheng, 1997; Griffiths & Tenenbaum, 2009; Lake et al., 2017; Zhao et al., 2022; Bramley et al., 2023).

In order to be compositional, human causal cognition must support representations that can easily be composed with each other, a bit like Lego blocks. This requirement implies that causal representations should ideally be *invariant* (or *modular*, or *disentangled*). Invariant causal representations adequately represent the structural relationship between two variables (or classes of variables), regardless of idiosyncratic facts about the system in which these variables are currently embedded (Cheng, 1997; Woodward, 2003, 2021; Sloman, 2005; Hiddleston, 2005; Bye et al., 2023; Schölkopf et al., 2021; Goyal & Bengio, 2022; Torresan & Baltieri, 2024). A representation that is not invariant (i.e. that only describes how the causal mechanism works in one particular context) does not adequately support compositional reasoning, because it cannot usefully be 'exported' outside of the current context.

The structural equation $E := 0$ is for example not an invariant representation of the relationship between E and P in the machine above. This is because the equation says that the state of E does not depend on P, but in fact there would be a dependence between the two variables if we plugged more nodes into the machine, for example if we added a generative input A:



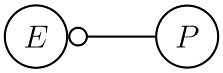
The equation $E := 0$ is not an invariant repreentation because it does not tell us to expect that the state of P might become relevant for the state of E once we add node A.

Integration functions

Good systems for causal representation tend to represent causal relationships as 'modules' that can be flexibly composed together. Modular representations are useful if we have a rule that tells us how composition works. This role is fulfilled by an *integration function* that specifies how the causal influence of various causes combine together to determine the value of a variable (Cheng, 1997; Waldmann, 2007; Lucas & Griffiths, 2010). Again, we can illustrate this notion in the context of our toy machine universe. The relevant integration function can be expressed in words as: ‘A node is activated if it receives at least one stimulation from a generative link, and no stimulation from a preventative link’. Assuming for simplicity that the machines are deterministic[4], we can specify the integration function formally as:

$$V := \max(G_1, \ldots, G_n)(1 - \max(P_1, \ldots, P_n))$$

Where G_i denotes a parent of V that is linked to V via a generative link, and P_i is a parent of V that is linked to V via a preventative link. The max() operator returns the largest value in a list of variable states; it evaluates to 0 if there is no variable within its scope. Consider our two-node machine from earlier:



In that system, the relationship between P and E is represented (graphically) by the preventative link going from P to E. This information can be passed to the integration function to determine that the state of E is given by:

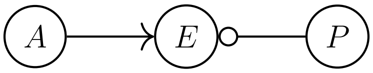
$$E := \max(0)(1 - \max(P))$$

[4] In stochastic settings, commonly-used integration functions are the ‘noisy-logical’ functions, such as Noisy-OR and noisy-AND-NOT, which take as input the ‘causal power’ parameters characterizing the link between two variables (Pearl, 1988; Cheng, 1997; Glymour, 1998; Yuille & Lu, 2007). Noisy-logical integration functions allow us to construct a Causal Bayes Net (CBN) describing a given causal system. In a CBN, the value of variable is not determined by a structural equation but by a conditional probability distribution (Pearl, 2000). Importantly, a Noisy-logical integration function, and the associated causal power parameters, are formally distinct from the CBN that they are used to construct. For example, the conditional probabilities making up the CBN are not typically the same numbers as the causal power parameters. So, while causal power parameters are invariant representations, CBNs are not invariant, for the same reason as SCMs.

which can further be simplified as:

$$E := 0$$

Considering now the machine with the added node A:



The state of E is there given by:

$$E := \max(A)(1 - \max(P))$$

$$E := A \& \neg P$$

In sum, when we represent a node P as being wired to another node E via a preventative link, this representation can be used as an input to the integration function (along with information about the other nodes wired into E), across many different possible machines. This is unlike a structural equation, which only contains information specific to a particular machine. Note that because the max() operator is agnostic about the number of inputs it takes, the integration function is flexible enough to represent any pattern of wiring, including the case where the node has no parents.[5] This flexibility of the integration function is crucial to the invariant nature of causal representations.

The integration function we gave above is just one example in the context of our toy machine universe. People use different integration functions in different domains, and these functions often have a different form, and are more complicated, than the one in our example (Waldmann, 2007; Lucas et al., 2014). Our example is meant to illustrate that good integration functions must have certain properties in order to enable the representation of invariant causal relationships—for example they should accommodate many possible different configurations of variables.

[5] Although we note that in some causal theories, integration functions might have non-optional arguments.

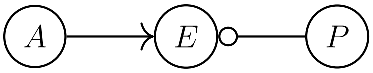
Empirically, research has established that people have systematic assumptions about invariant integration functions (Cheng, 1997; Waldmann, 2007; Griffiths & Tenenbaum, 2009; Woodward, 2021; Bye et al., 2023; Park et al., 2022; Cheng et al., 2022), that they can learn new integration functions on the basis of empirical data (Lucas & Griffiths, 2010; Lucas et al., 2014; Kemp et al., 2010; Kosoy et al., 2022; Jiang & Lucas, 2024), and that they can predict the behavior of novel systems in a compositional manner from their knowledge of the relevant causal laws (e.g. Battaglia et al., 2013; Baker et al., 2017; Zhao et al., 2022, 2024). Therefore, a theory of causal representation in human cognition should be able to emulate these capacities.

Causal models fail to fully capture invariance

As our earlier examples illustrate, structural causal models do not necessarily represent invariant causal relationships (Griffiths & Tenenbaum, 2007, 2009; Zhao et al., 2022).[6] A causal model is designed to describe how a particular causal system works, i.e. how a particular collection of variables are causally related to each other. The causal model contains enough information to compute what would happen in all possible states of the causal system, but this information might not generalize to other, related causal systems. In our toy universe of machines, a causal model describing a given machine might not allow us to predict the behaviour of another machine, even one constructed by making minimal modifications to the original machine.

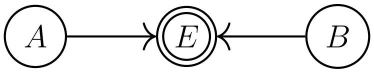
We saw one example of this problem earlier in this section. As another example, consider the fact that an SCM can collapse information about several causal mechanisms into the same structural equation. We can see this by contrasting the machine below with its structural equation representation:

[6] ‘Invariance’ can have slightly different meanings in the literature, and there is a sense in which causal models capture some of the invariance of causal relationships (see in particular Woodward, 2003). However, causal models do not fully capture invariance in the crucial sense defined above, which allows causal relationships to be exported from one context to the next.



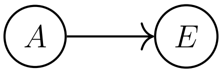
$$E := A\&\neg P$$

The machine is made of two causal mechanisms (the generative link between A and E, and the preventative link between P and E), but it is described by a single structural equation. To see why this equation discards information about the independent causal mechanisms, consider that the equation can also describe a very different machine:



In that machine, the double border around E indicates that it is a 'stubborn' node[7], which requires two stimulatory signals before it fires; the structural equation $E := A\&\neg P$ can be obtained simply by defining variable P as $P = \neg B$.

Because the structural equation can describe different possible machines, it cannot tell us what would happen if we disconnected the right-hand side node from one of the machines it describes: the correct answer is different for each machine, yet they are described by the same SCM. For the first machine, the system that results from taking away P is:

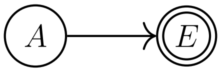


[7] Formally, accomodating stubborn nodes requires a slight generalization of the integration function described previously: we re-write this function as

$$V := \Sigma(G_1, \ldots, G_n)(1 - \max(P_1, \ldots, P_n)) \geq T$$

Where $\Sigma()$ evaluates to 0 if its scope is empty. With $T = 1$, we recover the integration function for a normal node, and when $T = 2$ we have the function for a stubborn node.

In this updated machine, E fires if and only A fires. In contrast, disconnecting B from the second machine yields the machine:



In that machine, E does not fire regardless of the state of A, since it is a stubborn node.

In sum, a structural equation is a good representation of the patterns of causal dependence that hold between a set of variables in a particular setting, but this representation is tethered to the idiosyncratic details of that setting; it cannot be usefully exported to other contexts.

Putting things back together.

We have just argued that causal models are not good representations of invariant causal relationships. Since people tend to represent invariant causal relationships, human causal cognition must be supported by representations that are richer than causal models like SCMs.

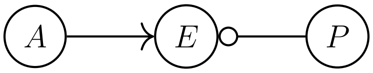
This claim is by no means novel (see e.g. Griffiths & Tenenbaum, 2007, 2009), but we think that its implications for the study of causal judgment have been under-appreciated (although see Maudlin, 2004; Hiddleston, 2005). In particular, it seems plausible that some causal judgments are readouts from invariant representations of causal relationships, rather than from causal models.

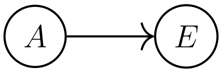
In the next section, we explore one way that this assumption could explain the distinction between productive and non-productive causation. We argue that if people construct invariant representations of causal relationships, they can use these representations to compute what would happen if one *disconnected* a variable from another variable. We then sketch a theory of productive causation that uses these variable-disconnection counterfactuals.

Toward a counterfactual account of productive causation

Variable disconnection

The assumption that people represent invariant causal laws makes it possible to consider a new kind of intervention, whereby we *disconnect* a variable from another. In our toy universe of machines, this operation simply consists of cutting the wire between two nodes; for example the machine below is shown both before and after disconnection of node P:





More formally, disconnecting a variable V from its child variable E consists in removing V from the inputs to the integration function for E.[8] For the machine shown above, the integration function says that the state of E before the disconnection is given by:

$$E := \max(A)(1 - \max(P))$$

$$E := A \& \neg P$$

After disconnection of P, we have:

$$E := \max(A)(1 - \max())$$

Which simplifies to:

$$E := A$$

[8] Note that variable-disconnection is not always a well-defined intervention. If we are using an integration function where some arguments are non-optional, then the outcome of some disconnections are not well-defined.

Figure 1 illustrates the difference between variable-disconnection interventions and the classical notion of intervention which changes the state of a variable (Pearl, 2000).

In our machine universe, our notion of disconnection can be easily visualized as removing the link between two nodes. But in other contexts, disconnection might be more abstract. Consider a physics simulation where two particles C and E exert a gravitational force on each other. To disconnect the causal influence that particle C exerts on particle E, we simply modify the equation that governs the behavior of particle E so that it does not represent the gravitational force from particle C anymore. The resulting situation is physically impossible, since particle C is still present in the scene but does not exert its force on particle E, without us having added any other object in the scene that would explain this. As such, the removal of the causal connection constitutes a 'miracle', in the same sense that standard state-change interventions in SCMs are miracles.[9] It is nonetheless easy to reason about the behavior of the particle after either kind of intervention, despite the physical impossibility of the overall situation.

Productive causation

Here we use the notion of variable disconnection to sketch a theory of the distinction between productive and non-productive causation. To motivate our definition, it will be useful to consider our double-prevention case again:

Double prevention: Suzy and Billy are playing in the garden. Suzy throws a rock at a nearby bottle. Billy is about to catch the rock, which would prevent it from breaking the bottle. However, Danielle pushed Billy away, preventing him from preventing the bottle from breaking. The bottle breaks.

Intuitively, the reason Danielle's action is not a productive cause is that Danielle's presence was not strictly speaking necessary for the bottle to break. This is because we

[9] Remember that in the standard notion of intervention, the value of a variable is set at a value chosen by the modeler, overriding the structural equation that describes the natural mechanism controlling the variable value (Pearl, 2000).

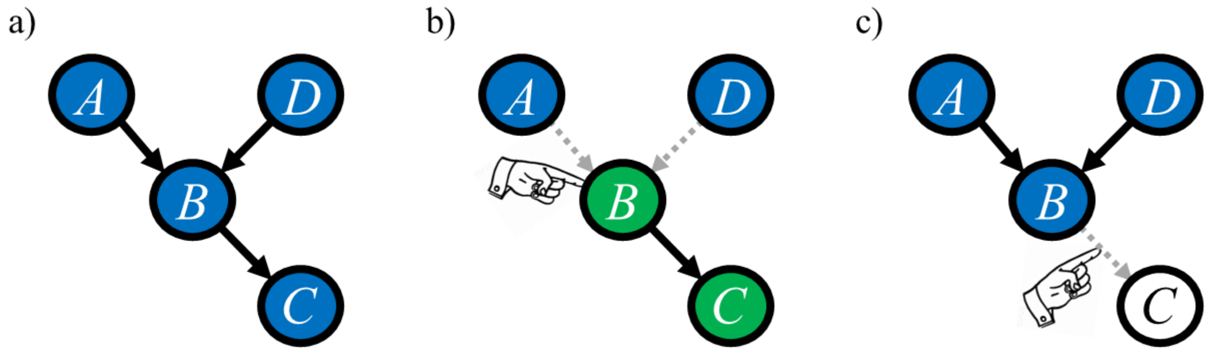


Figure 1

Difference between 'classical' state-change interventions (b) and variable-disconnection interventions (c). Panel (a) represents the initial state of a causal system, and dashed grey arrows represent disconnected causal links. b) To perform a state-change intervention on variable B, we first disconnect B from all its parents, and set B to a state of our choice (here, the green state). c) To perform a disconnection intervention on B, we disconnect the link between B and one of its children (here, C). The new state of the child variable is then determined by the relevant integration function. In contrast, the state-change intervention is possible even if we don't know the integration function for B, because the state of B is set manually by the modeler. Note that technically both types of interventions involve the disconnection of variables, but at different locations.

could have replicated the effect of Danielle's action by directly disconnecting Billy from the system instead. If Billy was not there in the first place, he would not have an opportunity to catch the rock, so the bottle would still have shattered. In contrast, Suzy's throw is a productive cause because the presence of Suzy was necessary for the outcome. If Suzy had not thrown her rock, the bottle would not have shattered; and we could not have made the bottle shatter by simply removing some variables from the system instead. We generalize and formalize these intuitions in our definition of productive causation:

Productive causation. If $C = c$ is an actual cause of $E = e$, and we cannot replicate the effect of $C = c$ on $E = e$ by instead disconnecting some variables from the system, then $C = c$ is a productive cause of $E = e$.

Our first requirement is for C to count as an *actual cause* of E. Roughly, event C being an actual cause of E means that C had some causal responsibility in bringing E about, but C does not need to be the 'main' cause of E. We do not commit to a particular theory of actual causation, but we agree with existing accounts that actual cause judgments require computing state-change counterfactuals (Halpern, 2016, see also Supplementary Information). The second requirement is that, after we have performed a state-change counterfactual showing that C is a cause of E, we cannot 'undo' the effect of that state-change counterfactual by disconnecting some variables from the system.

Our definition of productive causation is a preliminary one, and is relatively informal. We sketch a more formal definition in the Supplementary Information (along with examples of its application), but our goal is not to propose a definition that will survive all potential counter-examples—if the history of causal modeling is any guide, we suspect that doing so would be very difficult (Paul & Hall, 2013). Instead, in the next sections we will focus on experimental tests of the core idea underlying this definition, namely that intuitions about productive causation come from thinking about counterfactuals that disconnect a variable from the system.

Motivation for the experiments

Below we report a series of simple experiments designed to provide proof-of-principle support for our proposal. We ask participants to reason about simple 'machines', of the kind we introduced above. We predict that manipulating the counterfactual consequences of disconnecting a node from the machine will influence whether people construe a cause as a productive or non-productive cause. Importantly, this should be the case even holding constant: i) everything that happened in the actual world, ii) the causal model that describes the machine.

This prediction is at odds with standard counterfactual theories grounded in causal models: these theories predict that causal judgments should only track the counterfactuals that change the state of a variable. Our manipulation leaves intact the causal model describing a system, and therefore it does not affect the truth-value of state-change counterfactuals.

Because it is un-natural to explicitly ask participants about 'productive causation', we ask them to select the causal verb (CAUSE/ALLOW/ PREVENT) that better describes a given causal relationship (following previous research, e.g. Goldvarg & Johnson-Laird, 2001; Wolff & Song, 2003; Wolff et al., 2010; Kuhnmünch & Beller, 2005; Sloman et al., 2009; Cao et al., 2023; Beller & Gerstenberg, 2023; Thanawala & Erb, 2024). We make the assumption that when an event qualifies as a productive cause of an outcome, people are more likely to judge that the event CAUSED the outcome, relative to other verbs such as ALLOW. We leave a complete account of the semantics of these causal verbs for future research. Data and analysis code for all studies are available at https://osf.io/nsfx6/?view_only=3c9816af19ab463f9e9390740d622d86.

Study 1

We asked participants to make judgments about a simple machine in which a node A is wired onto a node E, see Figure 2. Node A can be in either of two states, and only one of them is associated with the activation of node E. We manipulate the consequences of the variable-disconnection counterfactual in which A is disconnected from E. Importantly, this manipulation does not affect the state-change counterfactuals, i.e. the consequences of changing the state of A in the full machine; see Figure 2 for illustration.

We implement our manipulation by allowing participants to observe what happens in a single-node machine where E is the only node. We call the state of E in such a machine its 'default state', df_E.

We predict that this minimal manipulation will influence the causal judgments that participants make about the full machine. In a situation where disconnecting A would change the value of E, participants should be more likely to say that the state of A Causes

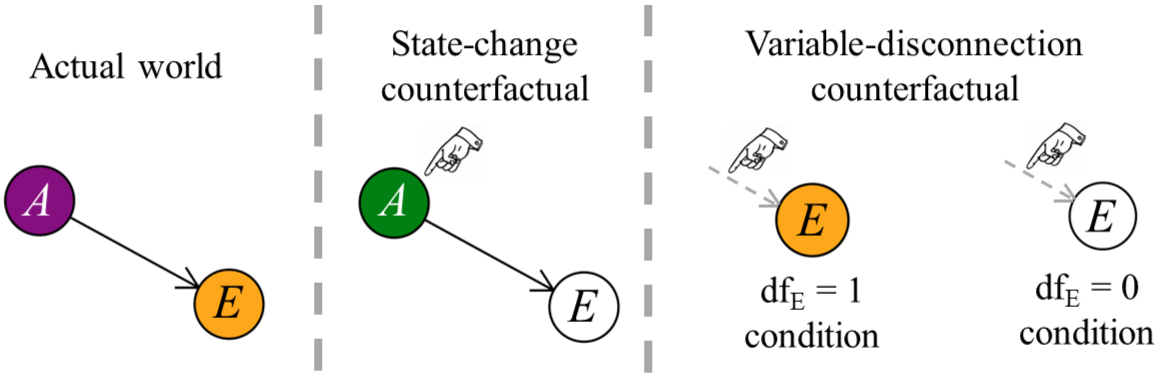

Actual world
A
E
State-change counterfactual
A
E
Variable-disconnection counterfactual
E
$df_E = 1$ condition
E
$df_E = 0$ condition

Figure 2

Schematic illustration of the causal dynamics in Study 1. Hand icons highlight the changes relative to the critical trial machine configuration, and were not shown to participants. In this example, E=1 in the actual world, but the experiment also had a condition where E=0 in the actual world (and the state-change counterfactual has E=1).

the state of E, compared to a situation where disconnecting A would not change the state of E. In other words, participants should be especially likely to say that the state of A Causes the state of E when the actual state of E does not match its default state.

This prediction follows from our definition of productive causation. Consider for example the situation depicted in Figure 2. First, in the the $df_E = 1$ condition, the state of E in the actual world matches its default state. A state-change counterfactual setting A to the green state would de-activate E. However, we could then re-activate E simply by disconnecting node A. Therefore A being purple is not a productive cause of E's activation. Consider now the $df_E = 0$ condition, in which the actual state of E does NOT match its default state. In this situation, a state-change intervention making A green de-activates E, and we cannot then re-activate E by disconnecting A from the machine. Therefore A being purple is a productive cause of E's activation.

In contrast, SCM-based approaches to causal judgment predict that our manipulation should have no effect, since our manipulation does not affect the truth-values of counterfactuals that change the states of variables.

Methods

Participants were introduced to a simple world of ‘machines’ composed of nodes that can be wired together. Each participant interacted with two different sets of machines: in one set, the default value of the effect node E was 0 ($df_E = 0$ condition), while in the other set, the default value of E was 1 ($df_E = 1$ condition). For each set of machines, participants first were able to learn the rules that govern the machines by observing different machines in different states (training phase); then they were asked to make causal judgments about different configurations of the full two-node machine (test phase).

The training phase consisted in a prediction task with feedback. Later in the test phase, participants were shown the full two-node machine, for which the state of both nodes was known. They were asked to select what was the best explanation for the state of the lower-most node. Participants had to select the explanation from a multiple-choice menu; each explanation related the state of the lower-most node (called E) to the state of the node above (node A). The task consisted in judging whether the state of node A stood in a [cause/allow/make no difference/prevent] relation to the state of node E.

In sum, our design had two within-subject manipulations. We manipulated the default value of E ($df_E = 0$ vs $df_E = 1$), and the actual-world value of E in the test trial ($E = 0$ vs $E = 1$).

Procedure

Participants first completed a consent form, and read a short set of instructions. Then they completed the main task (described below) twice, each time with a different set of machines (one where $df_E = 0$, another where $df_E = 1$). The sets were differentiated by the shape of their nodes: one set had diamond-shaped nodes, while the other set had circle-shaped nodes (shape assignment was randomized). We randomized whether the $df_E = 0$ or $df_E = 1$ condition was presented first.

The main task consists of a training phase and a test phase. In the training phase, participants completed 16 prediction trials. Each prediction trial displayed one machine, for which the state of all nodes was displayed except for the lower-most node, which was greyed and had an '?' sign inside. Participants were asked to predict whether that node was ON or OFF, by clicking on one of the two corresponding buttons. Clicking the button revealed the node's state: the node turned orange if it was ON, and white if it was OFF. The participant was also told whether their prediction was correct.

Prediction trials involved two different machines:

-a machine composed of only a single node (that we will call 'E'),

-a machine composed of two nodes: node A being wired into node E.

The two-node machine could be in two different states: node A could be in one of two states (a_1 or a_2, indicated by color), and node E could be ON or OFF (one state of A was associated with E being off, and the other state was associated with E being on). Machines were deterministic, such that for the same configuration the same prediction was always correct (participants were not told this explicitly). Therefore there were three types of trials: the single-node machine, and the two-node machine in two different states. Participants made 8 observations of the single-node machine, as well as 8 observations of the two-node machine (4 observations per state).

On each trial, the screen also displayed the current trial number, as well as the number of questions correctly predicted so far (see Figure 3a).

In the test phase, each trial displayed the two-node machine, and the state of each node was known. Participants were asked to select the explanation that they thought 'best describes what is happening', among the following four explanations:

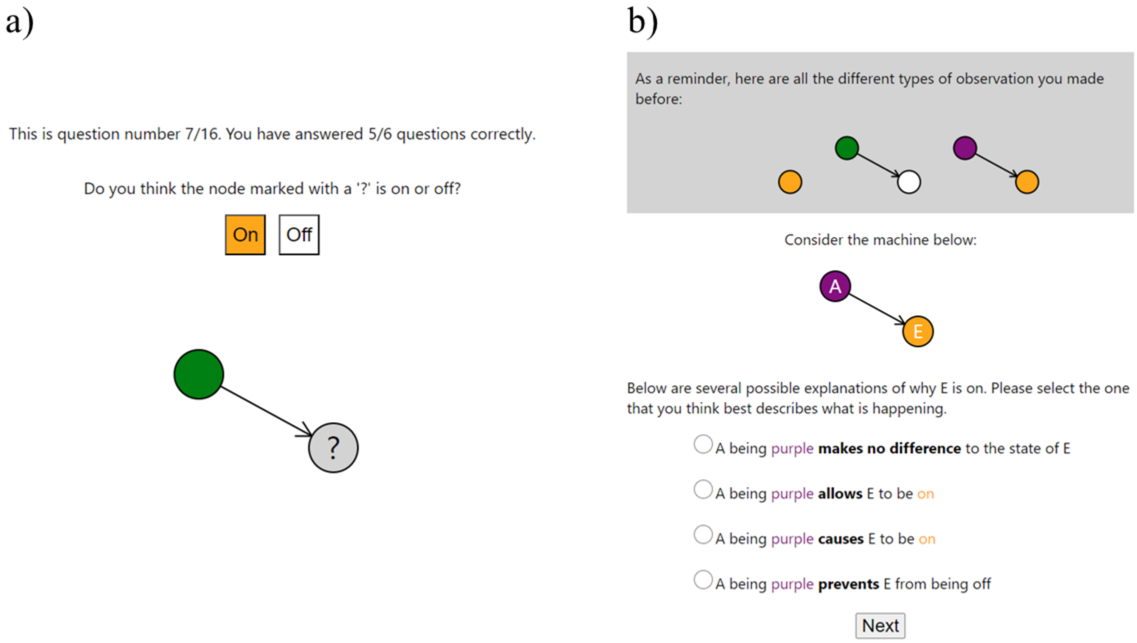
-A being [color] **causes** E to be [off/on]

-A being [color] **makes no difference** to the state of E

-A being [color] **allows** E to be [off/on]

-A being [color] **prevents** E from being [on/off]

Where [color] was node A's current state, and [off/on] was E's current state (or, for the 'prevent' statement, its opposite). The order in which these statements were presented



This is question number 7/16. You have answered 5/6 questions correctly.

Do you think the node marked with a '?' is on or off?

As a reminder, here are all the different types of observation you made before:

Consider the machine below:

Below are several possible explanations of why E is on. Please select the one that you think best describes what is happening.

○ A being purple **makes no difference** to the state of E

○ A being purple **allows** E to be on

○ A being purple **causes** E to be on

○ A being purple **prevents** E from being off

Figure 3

Screenshots of the experimental interface, Study 1. a) Training phase; b) Test phase.

onscreen was the same across all trials, but was randomized at the beginning of the study for each participant. Words shown in bold above were also bolded to participants, and [color] was displayed in a font of the corresponding color.

The two-node machine can be in two configurations: either node A is in state a_1, and node E is on, or node A is in state a_2, and node E is off. Therefore there were two test trials, one for each configuration (their order was randomized). The state of node A was indicated by two colors, for example the node might be black if in state a_1 and turquoise if in state a_2. Color assignment was randomized as follows. There were two pairs of colors: black/turquoise and green/purple. Which pair was assigned with which set of machines was randomized. Within each set, we also randomized which color was assigned to which state (except for node E, for which orange also meant On and white always meant Off).

During each test trial, the top of the screen displayed a summary of the observations the participant had made before. This reminder helped make salient the default value of the E node (see Figure 3b).

In between the two sets of machines, participants were told that the new machines they were about to see may obey different rules than the ones they had seen so far.

After participants completed the task for the two different sets of machines, they completed a short demographic questionnaire and were re-directed to Prolific for payment.

Participants

We recruited 79 US residents from Prolific (41 female, 2 other; mean age=40, SD=12). Participation was restricted to users with a 90%+ approval on the platform, who previously completed between 50 and 1000 studies. Participants were compensated GBP0.85, and median completion time was 7 minutes. In all experiments sample size was chosen to be similar to that used in recent experiments on causal verb selection with similar designs (Beller & Gerstenberg, 2023; Cao et al., 2023).

Results

Overall, participants seemed to learn the task, with 86% correct predictions in the last trial of the training phase. We excluded from subsequent analysis participants who did not perform significantly better than chance in the training phase, i.e. participants who got fewer than 65% trials correct (this is the threshold below which a binomial test results in a p-value larger than .05). We excluded 10 participants this way, resulting in a final sample of 69 participants.

Figure 4 displays the results for the test phase. Consistent with our hypothesis, we find that causal judgments are affected by whether the actual value of E matches its default value. In test trials where E's actual value is 0, proportions of Cause and Prevent choices are higher when E's default value is 1 ($df_E = 1$), compared to when $df_E = 0$. Proportions of Allow and Make No Difference follow the opposite pattern. When E's actual value is 1, most of these trends reverse: proportions of Cause choices are higher when E's default value is 0, while proportions of Allow and Make No Difference are higher when E's

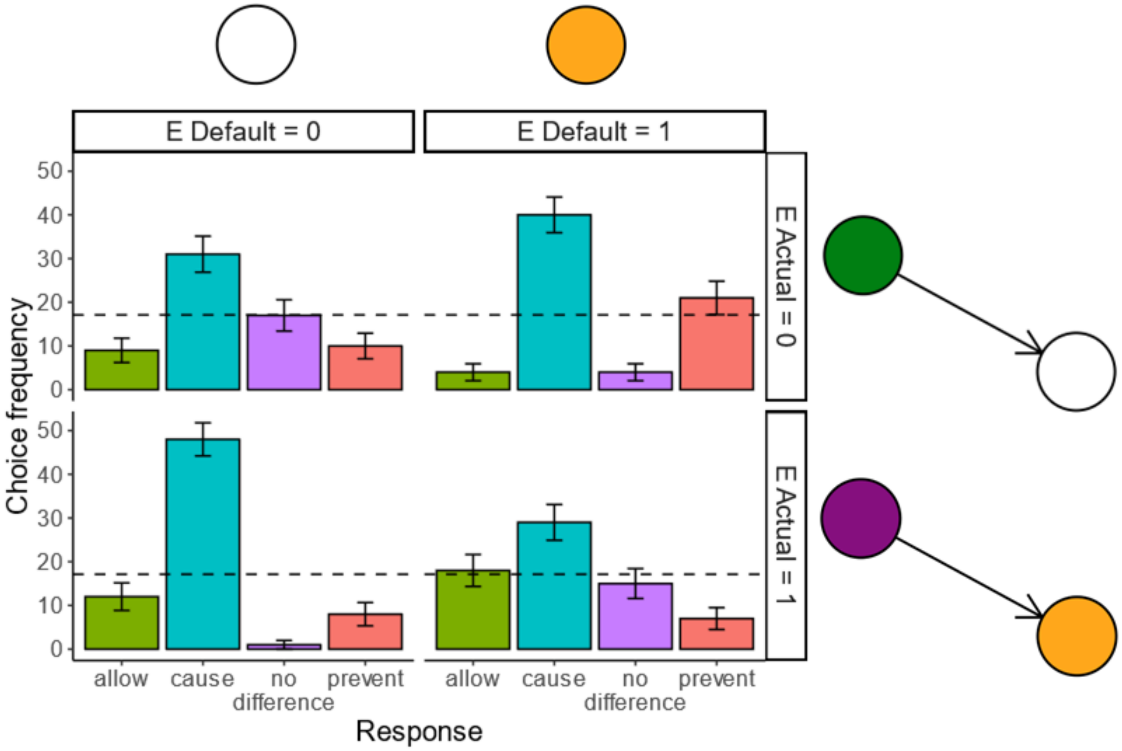


Figure 4

Frequency of responses across conditions, in the test phase, Study 1. Error bars represent standard errors. The dashed line represents expected frequency under random selection. The colors of the upstream node, and the shapes, are there for illustration, and were randomly assigned in the study.

default value is 1.

To formally evaluate these patterns, we performed a mixed-effects multinomial logistic regression using the brms package (Bürkner, 2017), with default priors. The model predicted participants' response from the actual value of E, the default value of E, as well as their interaction, and participant-level random intercepts. Using approximate leave-one-out cross-validation, we find that this model has a better fit than a null model without the interaction term (elpd=-264 vs elpd=-287, for the full and null model respectively).[10] In sum, whether the actual value of E matches its default value appears to meaningfully influence the relative proportion of responses.

To further unpack this result, we ran a series of mixed-effect logistic regressions, to assess what influenced the choice to select one response in particular, for example what influenced the decision to pick 'Allow' as opposed to the other three options. So, in the regression model for 'Allow', the outcome variable was a dummy variable with value 1 if the participant chose 'Allow' and 0 if the participant picked any other option. Again, the predictors were the actual and default value of E, as well of their interaction, and we used participant-level random intercepts.

We find that the interaction between E's actual and default value has a significant effect on the probability of selecting 'Cause', β = -2.66, $p < .001$, on the probability of selecting 'Allow', β = 1.69, $p = .041$, and on the probability of selecting 'Make No Difference', β = 5.80, $p < .001$. In contrast, it only had a marginally significant influence on the probability of selecting 'Prevent', β=-1.61, $p = .06$. Importantly, the sign of the interaction term is negative for Cause and Prevent, while it is positive for Allow and Make No Difference. Participants were more likely to select Cause or Prevent when E's actual value did not match its default value, but were more likely to select Allow or Make No Difference when E's actual value matched its default value.

One possible interpretation of the data is that participants selected Cause to a lesser extent in some conditions because they did not understand the relevant state-change counterfactuals in these conditions. For example, in conditions where E's actual state matches its default state, some participants might mistakenly think that changing the state of A would not change the state of E. In order to assess this proposal, in the Supplementary Information we analyze participants' responses as a function of their performance in the training phase (which indexes their ability to understand the

[10] We implement cross-validation using the loo function in brms, which uses Pareto-smoothed importance sampling (Vehtari et al., 2017).

consequences of changing the state of A). We find that the pattern of effects predicted by our theory are, if anything, more pronounced in participants who performed well in the training phase. Therefore, our results are unlikely to be due to a mis-representation of the state-change counterfactuals.

Supplementary Experiments

In the Supplementary Information, we report the results of two conceptual replications of Study 1 that address possible alternative interpretations for our results. In Supplementary Study 1b, we ask participants to reason about *contrastive* causal statements: instead of evaluating (e.g.) 'A being green causes E to be On', they evaluate 'The fact that A is green instead of purple causes E to be On'. This wording highlights the relevance of state-change counterfactuals, and therefore Study 1b constitutes an especially strong test of our claim that people also consider variable-disconnection counterfactuals when making causal judgments.[11]

Supplementary Study 1c addresses the possibility that our results are due to superficial patterns of co-variation. In Study 1, manipulating the default value of E also changes the on-screen co-variation between the state of A and the state of E: for example in Figure 3b (with $df_E = 1$), node E is Off only when node A is green, but if df_E was 0 then node A being green would co-vary less strongly with E being Off. Therefore, in Study 1c we let node A have three potential states, in a way that allows us to maintain co-variation constant while manipulating E's default value.

[11] We ran this replication experiment to rule out a possible explanation for the effect of our manipulation on participants' judgments in Study 1: some participants might have mis-interpreted the question as asking whether *the fact that node A was wired* caused the outcome. However, even in the original Study 1 this alternative explanation cannot account for all findings. In particular, our manipulation also had an effect on the ratio of Cause to Allow selections. When node A fails to prevent the activation of E ($df_E = 1$, $E = 1$), it seems strange to say that the fact that node A was wired into node E *allows* E to be ON. Therefore the mis-interpretation account would predict that rates of Allow selections should be highest in trials where disconnecting A changes the state of E, the opposite of what we actually find.

In both experiments we replicate the results of Study 1. Taken together, the data from the three experiments support the proposal that causal judgment is influenced by the consequences of variable-disconnection counterfactuals.

Discussion

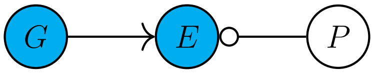
Under many counterfactual theories, causal judgments are derived from a causal model representation of the situation. This approach predicts that only counterfactuals that change the state of a variable should matter for causal judgment. The consequences of counterfactuals that disconnect a variable should have no influence of causal judgment, holding constant the state-change counterfactuals.

We tested this assumption by asking people to choose which of different causal expressions (Cause, Prevent, Allow, Make No Difference) best characterizes a given causal relationship. We find that manipulating the consequences of variable-disconnection counterfactuals did have an impact on people's choices, even holding constant the underlying state-change counterfactuals. People were more likely to describe an event as Causing an outcome when disconnecting the corresponding variable from the system would have changed the outcome; they were comparatively more likely to use Allow and Makes No Difference when disconnecting the variable would not have changed the outcome.

In Study 2, we attempt to conceptually replicate this finding in a more complex causal structure.

Study 2

In this Study, we consider a machine in which a generative node G and a preventative node P are wired onto the effect node E:



such that E is active if G is active but P is inactive. Intuitively, in a situation with $G = 1$, $P = 0$, it seems that G being active *caused* E to be active, while P being inactive *allowed* E to be active. We claim that this difference comes from the fact that people reason about variable-disconnection counterfactuals: E would still be active if P were disconnected, but not if G were disconnected.

However, this intuitive difference between the causal role of G and P might come from incidental aspects of the scenario, such as the fact that G is active while P is inactive — perhaps people treat absences as fundamentally different from positive events. Therefore, we design a minimal version of this scenario where G and P each have two symmetrical states, neither of which is explicitly labeled as active or inactive (although for the convenience of readers we will still use 0 and 1 to refer to the node states). We also refrain from giving explicit labels such as 'generative' and 'preventative' to participants. We only manipulate the consequences of variable-disconnection counterfactuals: in the state with $G = 1$ and $P = 0$, node E would still be active if we disconnected P, but not if we disconnected G; see Figure 5.

In the critical trials, we show participants a situation where $G = 1$, $P = 0$, and $E = 1$, and we ask them which causal verbs best describe the causal role of G and P. Importantly, both nodes play a symmetrical role in terms of state-change counterfactuals (intervening on either node would set E to 0, see top of Figure 5), such that an SCM-based account of causal judgment predicts they should be treated similarly.

We manipulate participants' belief about variable-disconnection counterfactuals by allowing them to observe a pair of two-node machines, in which only one of either G or P is wired into E, see Figure 5 (bottom). Participants can thus see that the presence of node G (when in the right state) is sufficient to activate E, while the presence of P isn't. We predict that this information will be sufficient to elicit a difference in judgments when people are asked about the causal roles of G and P in the full machine.

Procedure

The procedure was similar to Study 1, but because of the increased complexity we added an instruction phase at the beginning to help participants understand the causal

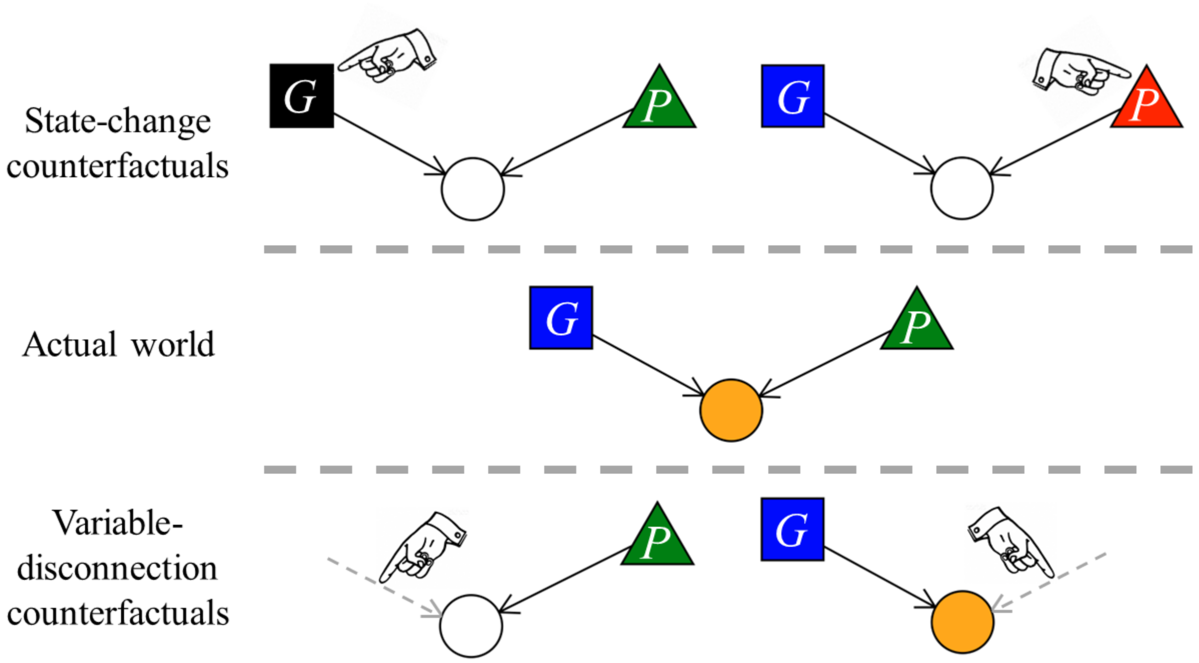

State-change counterfactuals
G
P
G
P
Actual world
G
P
Variable-disconnection counterfactuals
P
G

Figure 5

Schematic illustration of the causal dynamics in Study 2. Relative to the state of the machine in the actual world, state-change counterfactuals have the same consequences for G and for P, but variable-disconnection counterfactuals have different consequences: disconnecting G, but not P, changes the outcome. Hand icons highlight the changes relative to the critical trial machine configuration, and were not shown to participants. Letter assignment was randomized, and we used different letters than G and P.

structure. As such the experiment featured an instruction, a training and a test phase. During the training phase, participants were allowed to observe all 9 possible configurations of the system, while in the instruction and test phase, we let participants observe 6 different configurations (to avoid overwhelming them).

The 6 different machine configurations participants observed in the instruction and test phases were the following.

-A one-node machine with just E: its default value is 0.

-Two two-node machines: one with G wired into E and another with P wired into E. In the first machine, $G = 1$, $E = 1$; in the second machine $P = 0$, $E = 0$.

Finally, they observe the same three-node machine in three different states:

$G = 1$, $P = 0$, $E = 1$

$G = 1$, $P = 1$, $E = 0$

$G = 0$, $P = 0$, $E = 0$

During the instruction phase, these observations were presented on two screens. One of the screens showed all of the three machine configurations in which $G = 1$, while the other screen showed all three machine configurations in which $P = 0$. Each screen also showed the single-node machine. The order of presentation of machines on each screen was: first, the single-node machine, then the two-node machine, and then the two three-node machines (the three-node machines were presented in randomized order).

After this initial instruction phase, participants completed a training phase, similar to the training phase of Study 1. The training phase elicited predictions about all 9 possible machine configurations, including configurations that participants had not observed before (for example, a two-node machine with $G = 0$). Participants saw each machine configuration 3 times, for a total of 27 trials.

The test phase was similar to the test phase in Study 1. In the critical trials, participants were shown the three-node machine with $G = 1$, $P = 0$, $E = 1$. They were asked to make a causal judgment (choose the most appropriate causal verb) for both $G = 1$ and $P = 0$ (on separate pages—the order of presentation was randomized). The test phase also contained similar trials for all other three configurations of the three-node machine, but the critical trials were always presented first. Other trials were presented in randomized order. Each trial featured a reminder of the 6 observations presented in the instruction phase at the top of the screen.

Hypotheses and data exclusion criteria for this study were pre-registered (https://osf.io/a3prm/?view_only=208fc61e8e6644cf81eb513a95a69d06).

Participants

We recruited 100 US residents from Prolific (49 female, 2 other, mean age=40, SD=16). Eligibility criteria were the same as in previous studies, and participants were compensated GBP1.00 for participation. Median completion time was 8 minutes. Following our pre-registration, we excluded from analysis 15 participants whose proportion of correct answers on the prediction task was no better than chance ($n = 12$), or whose answers in the prediction task showed no positive correlation with the correct answers ($n = 8$).

Results

Participants gave the correct answer in the prediction task on 91.5% of trials, indicating that they overall understood the causal structure—this is significantly higher than chance performance of 50%, $p < .001$, binomial test. Participants also correctly generalized: on prediction trials featuring machine configurations that did not appear in the instruction phase, participants gave the correct answer 96.7% of the time (higher than chance, $p < .001$). This good generalization performance suggests that participants' inductive biases successfully guided them to our intended causal structure.

Figure 6 displays participants' judgments for the critical trials, i.e. their judgments for the machine configuration with $G = 1$, $P = 0$, $E = 1$. Participants described the causal role of the generative and the preventative node very differently. For the generative node, the modal answer was Cause, while for the preventative node the modal answer was Make No Difference, closely followed by Allow.

A mixed-effects multinomial logistic regression confirmed that participants made different choices for the generative and the preventative node. Using approximate leave-one-out cross-validation (loo), we find that a model including node type has a better fit than a null model that does not include node type as a predictor (elpd=-191 vs elpd=-221, for the full and null model respectively). Follow-up mixed-effects logistic regressions show that people selected Cause more often (β=2.38, $p < .001$), and selected Make No Difference less often (β=-3.49, $p < .001$) for the generative than the preventative

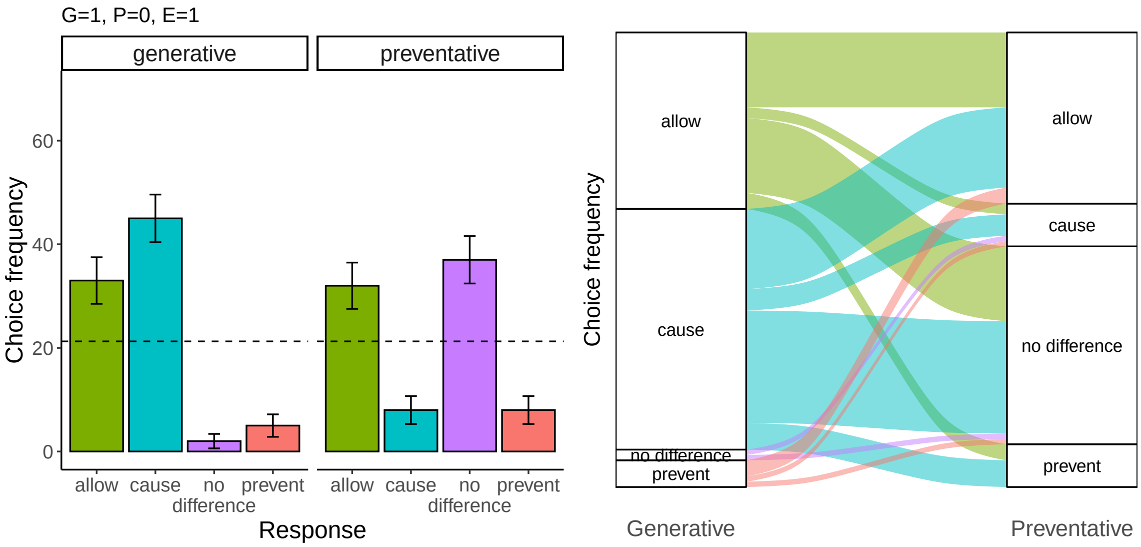


Figure 6

Study 2, Frequency of responses across conditions, machine configuration with $G = 1$, $P = 0$, $E = 1$. Error bars represent standard errors. The dashed line represents expected frequency under random selection. Both charts display the same data, in a different format; the chart on the right displays the frequency of participants making a given pair of selections.

node. There was no significant difference in the frequency of Allow and Prevent selections across node type (p=.87 and p=.39 respectively).

Interestingly (and contrary to our pre-registered predictions), the proportion of Allow statements did not change across condition. We find, however, that many of the participants who selected ‘Allow’ for the generative node selected ‘Make No Difference’ for the preventative node (n=14/33), while many of the participants who selected ‘Allow’ for the preventative node selected ‘Cause’ for the generative node (n=15/32). In other words, these ‘Allow’ judgments were not systematically made by the same participants (only 14 participants answered ‘Allow’ for both nodes), see Figure 6 right panel. We also find that the probability of selecting Allow is inversely related to performance in the training phase for the generative node, while it is positively related to training-phase performance for the preventative node, see Supplementary Information.

In the Supplementary Information, we also run additional analyses that suggest that misunderstanding of the causal structure is unlikely to explain our results.

Figure 7 shows the results for the other three machine configurations. In these configurations, the most interesting comparison for our purposes is between the role of the generative node in $G = 0$, $P = 0$, $E = 0$ and that of the preventive node in $G = 1$, $P = 1$, $E = 0$. These two situations are interesting because they are 'mirror images' of each other. In the first situation, E is counterfactually dependent on G (setting G to 1 would set E to 1), while it is counterfactually dependent on P in the second situation (setting P to 0 would set E to 1). A purely SCM-based approach would thus treat these nodes as playing an equivalent causal role. Instead we find that participants are more likely to say that the state of the preventative node prevented E (in $G = 1$, $P = 1$) than to say that the state of the generative node prevented E (in $G = 0$, $P = 0$).

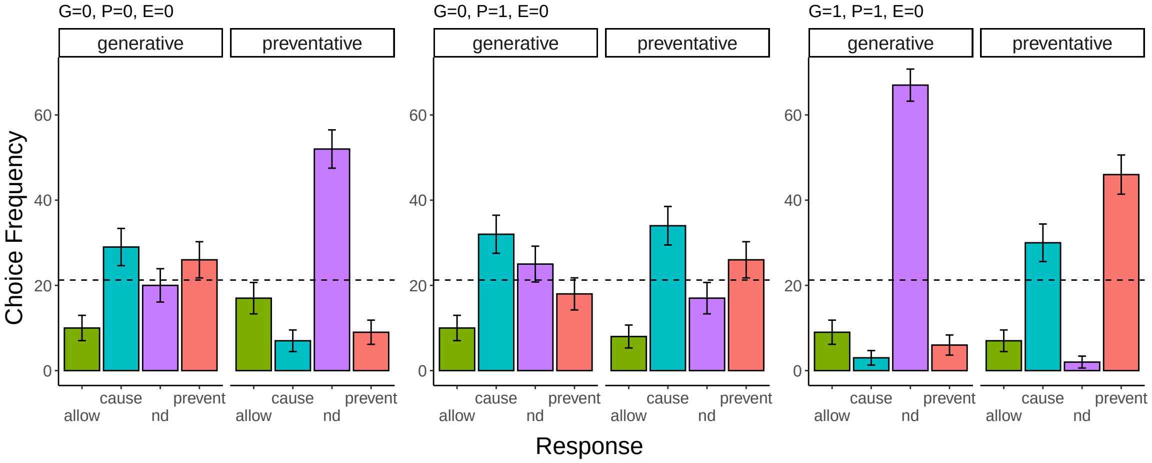


Figure 7

Study 2, Frequency of responses across conditions, non-critical trials. Error bars represent standard errors. The dashed line represents expected frequency under random selection. 'nd': 'make no difference'.

To formally assess this effect, we ran a mixed-effects multinomial logistic regression with the situation ($G = 0$, $P = 0$ vs $G = 1$, $P = 1$) as a predictor, and participant-level random intercepts; we only included judgments for the generative node for the first situation, and for the preventative node for the second situation. We find that this model has a higher fit to the data than an equivalent null model that does not include situation type as a predictor (elpd = -195 and elpd = -203 for the full and null model respectively).

Follow-up mixed-effects logistic regressions show that participants were more likely to select Prevent for the preventative node in the $G = 1$, $P = 1$ situation than for the generative node in the $G = 0$, $P = 0$ situation, β = 1.25, p = .002. Conversely, participants were more likely to select 'Make no difference' for the generative node in the $G = 0$, $P = 0$ situation, than for the preventative node in the $G = 1$, $P = 1$ situation, β = -2.55, $p < .001$.

Note that technically, neither node is a productive cause of $E = 0$ according to our definition (since 0 is the default state of E). Yet the two situations are not the same in terms of variable-disconnection counterfactuals, since disconnecting P changes the outcome in the $G = 1$, $P = 1$ situation, while disconnecting G does not change the outcome in the $G = 0$, $P = 0$ situation. This suggests that the mind makes fine-grained distinctions that go beyond the distinction between productive and non-productive causation, but that these distinctions could in principle be captured by a theory based on variable-disconnection counterfactuals.

The other comparisons are not as crucial for our hypothesis, for example within $G = 0$, $P = 0$, the variables G and P play an asymmetric role within the SCM (only an intervention on G would change the outcome). In $G = 0$, $P = 1$, the variables play a symmetric role in terms of standard counterfactuals, but they also have similar variable-disconnection counterfactuals (E remains inactive regardless of which node we disconnect). In that latter situation there was no evidence that node type influenced participants' judgments. A mixed-effects multinomial logistic regression with node type as a predictor had a slightly worse fit (elpd = -218) than a null model without node type (elpd = -214).

While our investigation is focused on variable-disconnection counterfactuals, the $G = 0$, $P = 0$ situation serves as a sanity check showing that participants are still also sensitive to state-change counterfactuals in our setting. In the $G = 0$, $P = 0$ situation, the variable-disconnection counterfactuals are the same (disconnecting either node leaves the outcome unchanged), but the state-change counterfactuals diverge (setting G to 1 activates E, while setting P to 1 does nothing). Accordingly, people are much more likely to say that the state of P made no difference to E, than to say that the state of G made no difference to E, $b = 2.05$, $p < .001$, mixed-effects logistic regression.

Discussion

Study 2 provides more evidence that variable-disconnection counterfactuals can influence people's causal judgments, even while holding constant the state-change counterfactuals.

We asked participants about an outcome caused by two other variables. One variable was a 'generative' variable which could bring about the outcome on its own, while the other variable was a 'preventative' variable which could only prevent the outcome. In a situation where the outcome happens, participants were most likely to say that the generative variable 'causes' the effect, and that the preventative variable 'allows' the effect or 'makes no difference'. Yet from the point of view of a Structural Causal Model, the two variables play a symmetrical role.

Since the state of the causal variables was denoted by colors (instead of the variables being ON or OFF), our finding cannot be explained by a tendency to ascribe more causality to variables that are present rather than absent. The effect cannot either be explained by the verbal labelling of the variables, since we never used terms like 'generative' or 'preventative' in the instructions given to participants. In sum, it seems that an SCM-based approach to causal language cannot explain the present results.

We also investigated situations where the outcome fails to occur. In one of the situations, the generative and preventative nodes are both inactive, while in the other situation, both nodes are active. From a structural equation perspective, the generative node (in the first situation) and the preventative node (in the second situation) play the same causal role. However, we find that participants are more likely to describe the preventative node than the generative node as preventing the outcome. Again, this result cannot be explained by differences in the labelling of the nodes. This finding provides a conceptual replication of similar findings by Walsh and Sloman (2011), in a much more minimal and controlled setting.

Study 3

In Study 3, we look at the reverse direction: We show participants a causal judgment describing a situation, and we see whether they make inferences about a counterfactual situation where a variable is disconnected from the system.

Intuitively, if you are told that Suzy caused the bottle to break, while Billy allowed the bottle to break, you might infer that removing Suzy from the situation would have prevented the breaking, whereas removing Billy from the situation might not have. Here we look at whether such inferences occur even with the minimal stimuli of our toy machine universe.

Specifically, we show participants a three-node machine, with nodes A and C both wired into node E. We show participants all 4 possible states of that machine. Nodes A and C can take either of two states (a_1 or a_2, c_1 or c_2); E is On whenever $A = a_1$ and $c = c_1$, and Off otherwise.

Crucially, for the state of the system where $E = 1$, we tell participants that $A = a_1$ ALLOWS E to be ON while $C = c_1$ CAUSES E to be ON. Then we ask participants to make predictions about reduced machines, constructed by disconnecting either A or C from the original machine.

We predict that in trials where $A = a_1$ or $C = c_1$, participants will be more likely to predict that E is On when it is connected to the C node than the A node. In other words, a node that was described as having caused the outcome is more likely to be predicted to lead to the outcome on its own, compared to a node that was described as having allowed the outcome.

Procedure.

Participants were first introduced to the basic setup of the machine universe, as in the previous experiments. We also told participants that different nodes can have different effects on an outcome, and for illustration we show them a node that can turn another node on, and a node that can turn another node off (these nodes have different colors and shapes than the nodes used in the main machine later in the study).

Then we showed participants the four possible states of the main machine, where nodes A and C are wired into node E. The nodes were not identified by letters to participants but had different shapes: A and C are a triangle and a square (counter-balanced), while E is a circle. Node states were denoted by colors, where a_1 and c_1 are blue and green (counter-balanced), while a_2 and c_2 are black and purple (counter-balanced). E is orange when it is On and white when it is Off. We also counter-balanced the position of A and C on the screen (left or right).

Then, participants completed a training task like in previous studies, where they made a binary prediction for the state of node E in each of the four possible states of the full machine, each presented four times (for a total of 16 trials, presented in randomized order).

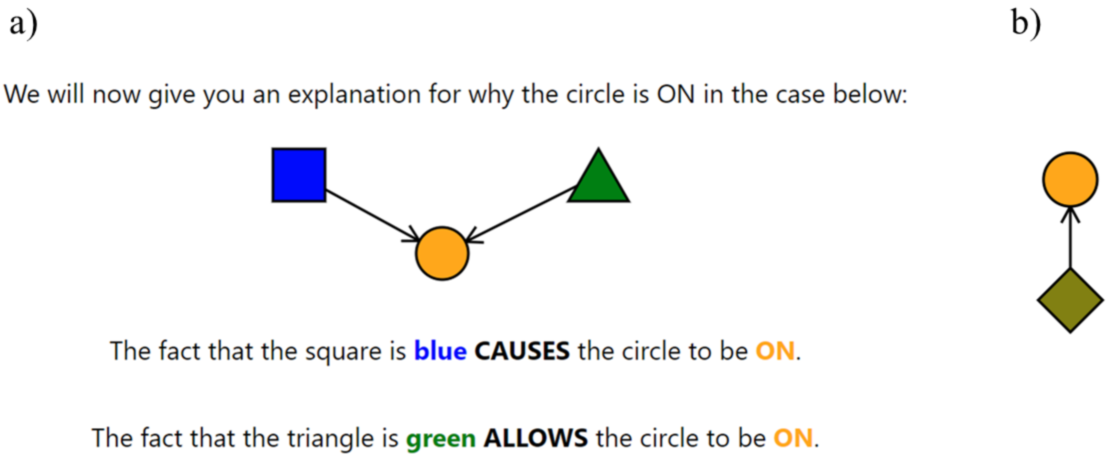
After the prediction task, an 'exposition' trial provided an explanation for why E is On in the configuration where $A = a_1$ and $C = c_1$ (see Figure 8a). We told participants (e.g.):

'The fact that the triangle is blue CAUSES the circle to be On,

The fact that the square is green ALLOWS the circle to be On.'

(the actual assignment of the shapes and colors was randomized).

This exposition trial was followed by the four main test trials, in which participants were shown a two-node machine that is the same as the full three-node machine they saw



We will now give you an explanation for why the circle is ON in the case below:

The fact that the square is **blue** **CAUSES** the circle to be **ON**.

The fact that the triangle is **green** **ALLOWS** the circle to be **ON**.

Figure 8

Study 3, partial screenshots of the experimental interface for a) the exposition trial, b) the introduction of the new node that precedes the far-transfer trials.

before, except with one of the nodes removed. That is, they saw A → E or C → E, depending on the trial. There were two trials per machine, one for each possible state of the upper node.

On each test trial, node E (the circle) was greyed, with a question mark at the center of the node (just as in the prediction trials). Participants were asked whether they think the circle is On or Off. Following this binary choice, they were asked to rate their confidence in their choice, on a 1-7 Likert scale from 'not at all' to 'very confident'. Each trial was presented on a separate page, and the order of the main prediction trials was randomized. On each trial, the top of the page displayed a reminder of the 'allow' and 'cause' judgments made in the exposition trial.

For exploratory purposes we also included two 'far-transfer' trials, where participants make predictions about a machine in which A or C are wired into E along with another, new node. Specifically, the page directly after the main test trials introduced another generative node, an olive diamond node connected to the circle, which is On (see Figure 8b). In the two 'far-transfer' trials that followed introduction of this new node, participants saw two three-node machines where either the A or C node, with state a_2 or c_2, have been added to this new machine. They again had to predict whether the circle is On or Off and rate their confidence.

The order of the far-transfer trials was randomized. At the top of the page, there was a reminder of the judgments made in the exposition trial, as well as a display of the N → E machine (N is the new diamond node). Then participants completed a short demographic questionnaire and were taken to Prolific for payment. Hypotheses, statistical transformations and data exclusion criteria for this study were pre-registered at https://osf.io/cya8k/?view_only=6fc28c92f2b5405bb5652d6232ee2fdf.

Participants

We recruited 95 US residents from Prolific (47 female, 2 other, mean age=39, SD=18). Eligibility criteria were the same as in previous studies, and participants were compensated GBP0.75 for participation. Median completion time was 6 minutes. Following our pre-registration, we excluded from analysis 11 participants whose performance in the training task was not significantly above chance, for a final sample of 84 participants.

Results

Participants gave the correct answer in the training task on 93.7% of trials —significantly higher than chance performance of 50%, $p < .001$, binomial test.

Figure 9 displays the results from the test phase. Following our pre-registration, participants' responses were coded on a scale from 1 to 14, where 1 represents full confidence in 'Off' and 14 represents full confidence in 'On'.[12] We call 'Allow trials' those

[12] Specifically, a rating from a participant predicting 'On' was coded as 7+confidence, while a rating from a participant predicting 'Off' was coded as 8-confidence.

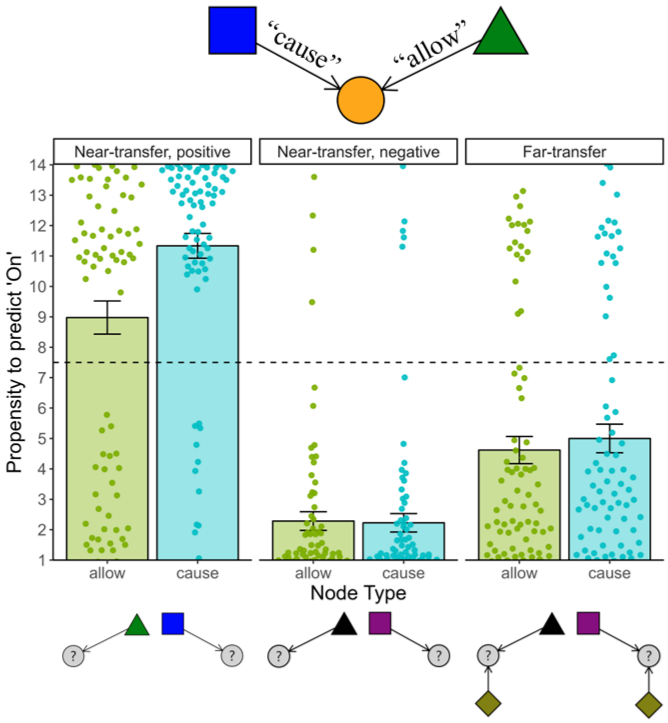


Figure 9

Propensity to predict 'On' on test trials, as a function of the causal verb used to describe the node from the original machine, Study 3. The exposition trial is shown at the top, the test trials at the bottom. Dots represent individual ratings. Ratings above the dashed line represent participants who predicted 'On', ratings below the dashed line are for participants who predicted 'Off'. Distance from the dashed line represent confidence in one's answer.

trials where the machine features node A, which was previously described as Allowing the outcome, and we call 'Cause trials' those trials where the machine features node C, previously described as Causing the outcome.

Consistent with our pre-registered prediction, when the nodes have the same value as in the exposition trial ($A = a_1$ or $C = c_1$), participants predicted that the circle node would be On to a greater extent in Cause trials (M=11.30, SD=3.73) relative to Allow trials (M=8.98, SD=4.98), V=185.5, $p < .001$, paired-samples Wilcoxon signed-rank test[13];

[13] Because our dependent variable is not normally distributed, we use a non-parametric test instead of a

see ‘Near-transfer, positive’ panel on Figure 9.

For robustness, we also compared the relative frequencies of ‘On’ and ‘Off’ predictions (ignoring confidence ratings) across these two conditions, finding that participants were more likely to predict On in Cause trials (N=71/84) than Allow trials (N=54/84), McNemar test, $\chi(1)$=23.9, $p < .001$. This result suggests that the causal verbs used to describe a situation influences the inference that people make about counterfactual situations where variables are disconnected from the system.

In contrast, there were no differences between Allow and Cause trials in either the ‘near-transfer, negative’ trials (machines C → E and A → E with $C = c_2$, $A = a_2$), V=233, $p = .74$, or the ‘far-transfer’ trials, V=439, $p = .52$.[14]

Since Study 3 is a ‘mirror’ version of Study 2, we can also think of its results in relation to the results of Study 2. In that study, speakers were much more likely to use Cause to describe the generative compared to the preventative node, while they were equally likely to use Allow to describe the generative and the preventative node. A listener that was trying to ‘invert’ an accurate model of the speakers in Study 2 would thus infer (in the current study) that the Cause node is likely to lead to the outcome on its own, while the Allow node may or may not do so. This pattern is similar to what listeners inferred in the current study.

t-test, in a slight deviation from our pre-registration; using paired-sampled t-tests instead yields identical results.

[14] We pre-registered as an exploratory hypothesis that on the ‘far-transfer’ trials, participants might give higher ratings in the Cause relative to the Allow condition. The rationale for that prediction was that participants might have made abstract inferences on the basis of the causal judgments, classifying the Allow node as more likely to exert ‘inhibitory’ causal power when in its negative state. Speculatively, the fact that we find no difference in the far-transfer trials might be explained by the fact that participants did not make these more abstract inferences, or that integrating these inferences with information about a new node would have been too cognitively demanding.

Discussion

Results of Study 3 suggest that listeners can use the causal language that describes a causal relationship to make inferences about counterfactual situations in which a variable is disconnected from the causal system. In a situation where two events jointly cause an outcome, and the events are respectively described as Causing and Allowing the outcome, participants made different inferences about what would happen if the corresponding variable was disconnected from the system. Participants were more likely to predict that the outcome would still occur if the variable corresponding to the Cause event was still present while the variable corresponding to the Allow event was disconnected, relative to vice-versa. Accounts of causal language based on SCMs cannot predict this result.

General Discussion

Humans are remarkable in their capacity to generalize what they learn. The flexibility of causal cognition suggests that the mind strives to construct invariant representations of causal mechanisms, that can easily be re-used across contexts.

With this hypothesis in mind, we can ask what representations people use when they make a causal judgment about a particular system. Many existing theories implicitly assume that people use a relatively impoverished representation: a Structural Causal Model (SCM). An SCM designed to represent a particular system can discard a lot of information about the more general causal laws that explain why the system works the way it does. We suggest that causal judgments might instead be derived from a representation of these more general causal laws.

This hypothesis can explain why people make an intuitive distinction between ‘productive’ and non-productive causation, a phenomenon that counterfactual theories typically struggle to explain. We argue that causal models are simply not expressive enough to model this distinction. Our claim is supported by a series of simple experiments, in which we successfully influence people’s causal judgments by manipulating the consequences of ‘variable-disconnection’ counterfactuals, a manipulation that leaves intact the underlying causal model representation of the situation. In this section, explore the relationship between our account and other theories of causal representation, and discuss outstanding questions.

Causal model theory

Sloman et al. (2009) have argued that the meaning of CAUSE, ALLOW and PREVENT could be formalized in terms of SCMs (see also Cao et al., 2023).[15] According to their causal model theory of causal meaning, the appropriateness of a causal verb depends on the structural equation with which people represent the corresponding causal relationship. For example, the claim that C causes E is appropriate when the structural equation for E is of the form:

$$E := C$$

People are predicted to judge that C allows E when the structural equation for E takes the form:

$$E := C \& X$$

Where X is another variable (or set of other variables).

The results of our experiments suggest that this theory cannot fully account for the meaning of CAUSE and ALLOW: we were able to systematically influence the proportion of CAUSE and ALLOW statements while leaving the relevant structural equations intact.

In the Supplementary Information, we address the possibility that participants might have constructed more complex SCMs than we have been suggesting (for example, incorporating variables that represent whether a node is wired). We argue that even on this assumption, SCM-based theories do not seem to easily explain our data.

[15] Technically, Sloman et al.'s theory is meant to account for the representation of generic causal claims (e.g. Heat causes fires) instead of singular causation (e.g. the heat caused the fire). However, it could plausibly be extended to cases of singular causation. Sloman et al. also use ENABLE where we use ALLOW; here we abstract over this difference, based on the fact that these verbs seem to be very close in concept space (Wolff & Song, 2003).

Mental model theory

According to Mental Model theory, the meaning of causal expressions is determined by particular patterns of possibilities in the reasoner's mental model of the situation, along with a general assumption that causes precede their effects (Goldvarg & Johnson-Laird, 2001; Khemlani et al., 2014). The expression 'A allows E' means for example that A is necessary for E, i.e. that there is no possibility (in the relevant mental model) in which E occurs in the absence of A. In contrast, 'A causes E' means that A is sufficient for E: E occurs whenever A occurs. The theory also gives a definition of PREVENT, and makes predictions about omissive causation (Khemlani et al., 2018).

In order to derive predictions from Mental Model theory in the context of our experiments, we would need an account of how people create mental models for a given machine; i.e. an account of what possibilities people spontaneously represent. Consider for example the simple machine $C \to E$. Do reasoners include different possibilities in their mental model of the machine, depending on what would happen if we disconnected node C? As far as we know, existing implementations of the theory are silent on these questions. Developing such an account is a valuable direction for future research, although it is outside the scope of the present paper.

Process theories

Intuitively, it seems that intuitions about productive causation might be explainable in terms of the transmission of some physical quantity, like force or energy: Suzy's throw is a productive cause of the bottle breaking because one can trace a physical process all the way from Suzy's throw to the bottle breaking. Several philosophical accounts of causation build on this insight (Dowe, 1992; Salmon, 1994).

In psychology and linguistics, the force-dynamics framework postulates that when people think about causation, they rely on representations akin to the ones found in Newtonian mechanics: force vectors that additively combine to affect the motion of an entity in (physical or metaphorical) space (Talmy, 1988; Wolff, 2007; Copley & Wolff, 2014). Force dynamics give an intuitively compelling, and empirically successful, account of causal concepts like PREVENT, CAUSE and ALLOW (Wolff, 2007; Wolff et al., 2010; Wolff & Barbey, 2015).

According to the theory, causal judgments are derived from vector addition on the represented forces. Consider for example our rock-throwing example. The bottle has a tendency to remain intact, which is mentally represented as a force (the 'patient force' P). This force points in a different direction than the endstate E, i.e. the bottle being broken. By throwing the rock, Suzy exerts an 'affector' force A on the bottle. By performing vector addition on the forces exerted by A and P, people can compute the endstate E, namely that the bottle will break. Wolff holds that people are inclined to use CAUSE to describe situations where force A points in a different direction than the patient force P, and P points in a different direction than E. The force dynamics treatment of ALLOW is slightly more complicated. Wolff et al. (2010) argue that judgments of ALLOW result from a hybrid process that combine force representations with counterfactual thinking: people represent the Affector force as the resultant of two forces in a double prevention case (the force exerted by the preventer, and the force exerted by the double preventer).

The force dynamics framework successfully accounts for many patterns of causal intuitions. At the same time, it carries a strong commitment about representation, namely that people maintain vector representations of the relevant factors at play in a scene (even when thinking about non-physical domains), and perform vector addition to make causal judgments.

Our framework can help explain why force dynamics theory is often a successful account of causal judgment, while offering a way to explain the same phenomena with much leaner representational commitments. We submit that the force theory is successful because force representations are invariant causal representations. For example, in Newtonian mechanics, the force that an object C exerts on another object E does not depend on the other forces exerted on E. As such, forces compose in a predictable way, and we can use a force representation to compute what would happen if we removed or added other forces from an interaction. In other words, when we represent a situation in terms of Newtonian-like forces, it is possible to compute the consequences of variable-disconnection counterfactuals. We submit that this is why force dynamics can account for the difference between cases of productive and non-productive causation.

Of course, vector representations of forces are only a small subset of the set of invariant causal representations. So, in order to model people's causal judgments in non-physical domains, we may not need to assume that they perform vector addition over force-like representations, as the force dynamics framework does. Instead all we need to do is posit that people use representations that are sufficient invariant to support variable-disconnection counterfactuals.

The Counterfactual Simulation Model

Gerstenberg and colleagues have argued that people make causal judgments by running counterfactual simulations over probabilistic programs: for example, they make causal judgments about physical events by using an internal physics engine (which approximates Newtonian dynamics; Gerstenberg et al., 2021). One key assumption of the 'Counterfactual Simulation Model' (CSM) is that when people make causal judgments, they assess both 'whether-causation' (checking whether the outcome E would have occurred in the absence of candidate cause C), and 'how-causation' (checking whether E would have occurred in a different way if C had occurred in a different way). The counterfactual interventions involved in assessing how-causation consist in slightly changing the way that C happened—for example, in a physical setting it can involve slightly modifying the velocity or trajectory of a billiard ball.

Beller and Gerstenberg (2023) have developed a theory of causal meaning that can capture how people use causal verbs in a simple physical setting. They argue that people are more likely to describe C as CAUSING E when the causal relationship exhibits both how-causation and whether-causation, while they are more likely to describe C as

ALLOWING E when the relationship exhibits only whether-causation.

Our framework can give a high-level explanation for the fruitfulness of the CSM approach to causal language. Beller and Gerstenberg (2023) assume that people represent physical interactions using an internal approximation of Newtonian mechanics; as we argue in the previous section, such a representation is in principle expressive enough to support compositional causal reasoning, and therefore to compute variable-disconnection counterfactuals.

On the other hand, our experimental data suggest that the human mind distinguishes between different kinds of causation even in the absence of the cues that the CSM highlights. In our experiments, all variables had discrete values. As such there were no-difference in 'how-causation' between events that participants judged as causing and events that participants judged as allowing an outcome.

Similarly, there were no explicit mention of vector-like forces in our experiments, which argues against a force dynamics account. Nonetheless, proponents of these approaches (CSM or force dynamics) might argue that participants in our experiments were spontaneously inferring the presence of forces, or that they represented the binary states of the nodes in a continuous format. For example, participants might have conceived of the nodes as transmitting an electrical force to their children, and might have based their judgments on this force representation. Therefore, more empirical research is needed to fully arbitrate between our approach and these other theories.

Other counterfactual approaches

Counterfactual theorists could argue that the distinction between productive and non-productive causation arises from a *quantitative* difference in the strength of our causal intuitions, instead of being a *qualitative* distinction. Human causal judgments are typically graded (Morris et al., 2019; O'Neill et al., 2022), so people might simply view Suzy as 'more of a cause' of the bottle breaking than Billy. Supporting this view, philosopher Paul Henne and his collaborators have shown that computational models designed to explain gradation in causal judgment (Icard et al., 2017; Quillien, 2020) can account for many intuitions about productive causation (Henne et al., 2017, 2019; Henne & O'Neill, 2022; O'Neill et al., 2022). In a double prevention case, for example, the productive cause is typically the variable that covaries most highly with the outcome, across possible counterfactual states of the system (O'Neill et al., 2022).

We are sympathetic to this perspective, but we believe it can only explain some of the relevant data. Notably, experimental manipulations that have an influence on graded causal intuitions often leave production intuitions unaffected. In a recent study, Thanawala and Erb (2024) manipulated whether the agents in a double prevention case acted intentionally. In the situation where only the double-preventer acted intentionally, participants gave higher causal ratings to the double-preventer (e.g. Danielle preventing Billy from catching the rock) than the producer (e.g. Suzy throwing the rock), when asked whether the outcome happened because of that agent. Yet in the same situation participants still described the double-preventer as Allowing the outcome, while they described the producer as Causing the outcome. Similarly, it is difficult to get participants to agree that the double-preventer *broke* the bottle, even when they strongly agree that the bottle broke *because* of the double-preventer (Rose et al., 2021).

Other counterfactual theorists suggest that SCMs might account for people's causal judgments if we also assume that people hold that variables have 'default' and 'deviant' values (Maudlin, 2004; Menzies, 2004; Hall, 2007; Hitchcock, 2007; Gallow, 2021, but see Wysocki, 2023). The default value of a variable is (roughly) the value it has when nothing else is acting on it (Hall, 2007). With this additional representational baggage, SCM-based theories can give reasonable accounts of actual causation, and of the distinction between productive and non-productive causation (see especially Hitchcock, 2007).

While this work is insightful, theorists have added the default/deviant distinction to the SCM framework in a relatively post-hoc way. In our framework, the idea that variables have something like a default value emerges naturally, as a byproduct of compositional causal reasoning. If a reasoner has an invariant representation of the causal laws that determine the value of a variable, this representation will usually specify what happens in cases where no other variable is exerting a causal influence on it (otherwise, the representation would not be sufficiently general). One can view that particular value as the 'default value' of a variable.

As such, we submit that our framework sheds light on why the default/deviant distinction has been useful in past causal modeling work: information about default values indirectly encodes some information about the invariant causal laws that govern the system.

Heterogeneity in participants' judgments

Our empirical data shows a lot of variation in people's judgments: within the same condition participants can be split almost equally between two or three different options. Participants may think that in some situations several words apply, for example that it is adequate to both say that an event C caused and allowed an event E to happen. As evidence for this conjecture, in studies where participants give a separate rating for each causal expression, there are situations where most participants select both Cause and Allow to describe a causal relationship (Cao et al., 2023).

We assume that people have a tendency to use Cause (relative to Allow) more frequently to describe productive causes, but this tendency appears to be a soft preference. In particular, in Study 1 Cause was the modal answer in all conditions, although its relative prevalence compared to other options changed in the direction predicted by our theory. We speculate that the relatively abstract nature of our stimuli contributed to this tendency. In order to test our theory in a controlled setting, we used manipulations that are quite minimal compared to most experiments on causal meaning. Participants who did not have vivid intuitions might have seen the word Cause as the 'safe' option, since it is arguably the most generic term, across the options we offered, that conveys the fact that an event exerts some causal influence on the outcome.

Toward a formal account of productive causation

We claim that our formal notion of variable-disconnection can help make sense of the intuitive distinction between productive and non-productive causes, and we have sketched a formal theory of productive causation along these lines. Much work remains to be done to refine the theory, and empirically test its more fine-grained predictions.

In doing so, it will be important to incorporate the role of time. Following much of the tradition in causal modeling, here we deliberately ignored temporal dynamics. This is an over-simplification, and there are reasons to think that intuitions about productive causation depend on how events unfold in time. As such, a complete formal account of these intuitions would need to incorporate a formal representation of time (see e.g. Nadathur & Lauer, 2020).

For example, Skow (2023) points out that temporal dynamics matter to the way people interpret cases of double prevention. Consider a case where striking down a pillar makes the roof fall down, by preventing the pillar from preventing the fall. It seems that striking down the pillar caused the roof to fall down (instead of merely allowing it; Schaffer, 2000). Skow suggests that we view double-preventers as productive causes when the preventer was already preventing the outcome before the double preventer intervened. A complete theory of productive causation would ideally explain why.

Conclusion

In causal cognition research, a lot of progress has been made by thinking carefully about the format of the representations that support causal thought. In this spirit, we suggest that structural causal models are not a sufficiently expressive representation to support the full wealth of causal judgments that people make. In turn, this argument suggests that one particular implementation of the counterfactual theory of causation has limitations when it comes to modeling causal judgment.

Ultimately, however, our work supports the counterfactual approach to causation, and the associated framework of interventionism (Pearl, 2000; Woodward, 2003). Our point is that the limitations of structural causal models do not necessarily mean that counterfactuals are inadequate for modeling causal judgment. Our proposed approach to causal judgment fully embraces the notion that people make counterfactual interventions on their internal representation of the world: we simply suggest that these representations aim to capture invariant causal relationships, instead of the idiosyncratic details of one particular causal system.

References

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Beckers, S., & Vennekens, J. (2018). A principled approach to defining actual causation. *Synthese*, *195*(2), 835–862.

Beebee, H. (2004). Causing and nothingness. In *Causation and counterfactuals*.

Beller, A., & Gerstenberg, T. (2023). A counterfactual simulation model of causal language.

Blanchard, T., & Schaffer, J. (2017). Cause without default. In *Making a difference* (pp. 175–214). Oxford University Press.

Bramley, N. R., Zhao, B., Quillien, T., & Lucas, C. G. (2023). Local search and the evolution of world models. *Topics in Cognitive Science*.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Bye, J. K., Chuang, P.-J., & Cheng, P. W. (2023). How do humans want causes to combine their effects? the role of analytically-defined causal invariance for generalizable causal knowledge. *Cognition*, *230*, 105303.

Cao, A., Geiger, A., Kreiss, E., Icard, T., & Gerstenberg, T. (2023). A semantics for causing, enabling, and preventing verbs using structural causal models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*(45).

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

Cheng, P. W., Sandhofer, C. M., & Liljeholm, M. (2022). Analytic causal knowledge for constructing useable empirical causal knowledge: Two experiments on pre-schoolers. *Cognitive science*, *46*(5), e13137.

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

Copley, B., & Wolff, P. (2014). Theories of causation should inform linguistic theory and vice versa. In *Causation in grammatical structures* (p. 57, Vol. 11). Oxford University Press Oxford.

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

Gallow, D. (2021). A model-invariant theory of causation. *Philosophical Review*, *130*(1), 45–96.

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.

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

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and machines*, *8*(1), 39–60.

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

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.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, *478*(2266), 20210068.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In *Causal learning: Psychology, philosophy, and computation* (pp. 323–345).

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

Hall, N. (2004). Two concepts of causation. In *Causation and counterfactuals* (pp. 225–276). The MIT Press.

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

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

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*.

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

Hiddleston, E. (2005). Causal powers. *The British journal for the philosophy of science*, *56*(1), 27–59.

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

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *The Philosophical Review*, *116*(4), 495–532.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, *62*, 135–163.

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

Jiang, C., & Lucas, C. G. (2024). Actively learning to learn causal relationships. *Computational Brain & Behavior*, 1–26.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243.

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

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

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

Konuk, C., Goodale, M. E., Quillien, T., & Mascarenhas, S. (2023). Plural causes in causal judgment. *Proceedings of the annual meeting of the cognitive science society*, *45*(45).

Kosoy, E., Liu, A., Collins, J. L., Chan, D., Hamrick, J. B., Ke, N. R., Huang, S., Kaufmann, B., Canny, J., & Gopnik, A. (2022). Learning causal overhypotheses through exploration in children and computational models. *Conference on Causal Learning and Reasoning*, 390–406.

Kuhnmünch, G., & Beller, S. (2005). Distinguishing between causes and enabling conditions—through mental models or linguistic cues? *Cognitive Science*, *29*(6), 1077–1090.

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*, e253.

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

Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, *31*(1), 107–127.

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

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

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

Maudlin, T. (2004). Causation, counterfactuals, and the third factor. In *Causation and counterfactuals*.

McGrath, S. (2003). Causation and the making/allowing distinction. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *114*(1/2), 81–106.

Menzies, P. (2004). Causal models, token causation, and processes. *Philosophy of Science*, *71*(5), 820–832.

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

Nadathur, P., & Lauer, S. (2020). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: a journal of general linguistics*, *5*(1).

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., Quillien, T., & Henne, P. (2022). A counterfactual model of causal judgment in double prevention. *Conference in computational cognitive neuroscience*.

Park, J., McGillivray, S., Bye, J. K., & Cheng, P. W. (2022). Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions. *Cognitive Psychology*, *132*, 101432.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan kaufmann.

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

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Quillien, T. (2020). When do we think that X caused Y? *Cognition*, *205*. https://doi.org/10.1016/j.cognition.2020.104410

Quillien, T., & Barlev, M. (2022). Causal judgment in the wild: Evidence from the 2020 u.s. presidential election. *Cognitive Science*, *56*(2). https://doi.org/10.1111/cogs.13101

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*. https://doi.org/10.1037/rev0000428

Rehder, B. (2017). Concepts as causal models: Categorization. In *The oxford handbook of causal reasoning* (pp. 347–376). Oxford University Press New York, NY.

Rose, D., Sievers, E., & Nichols, S. (2021). Cause and burn. *Cognition*.

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

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

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634.

Skow, B. (2023). Two concepts of double prevention. *Ergo an Open Access Journal of Philosophy*, *9*.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

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

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search.* MIT press.

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

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. *Causal learning: Psychology, philosophy, and computation*, 301–322.

Thanawala, H., & Erb, C. D. (2024). Revisiting causal pluralism: Intention, process, and dependency in cases of double prevention. *Cognition*, *248*, 105786.

Torresan, F., & Baltieri, M. (2024). Disentangled representations for causal cognition. *Physics of Life Reviews.*

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, *27*, 1413–1432.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive science*, *31*(2), 233–256.

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

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

Wolff, P., & Barbey, A. K. (2015). Causal reasoning with forces. *Frontiers in human neuroscience*, *9*, 1.

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

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive psychology*, *47*(3), 276–332.

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

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

Wysocki, T. (2023). Conjoined cases. *Synthese*, *201*(6), 197.

Xiang, Y., Landy, J., Cushman, F. A., Vélez, N., & Gershman, S. J. (2023). Actual and counterfactual effort contribute to responsibility attributions in collaborative tasks. *Cognition*, *241*, 105609.

Yuille, A. L., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. *Advances in neural information processing systems*, *20*.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? a non-parametric bayesian account. *Computational Brain & Behavior*, *5*(1), 22–44.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2024). A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour*, *8*(1), 125–136.