A Model-Invariant Theory of Causation

J. Dmitri Gallow
Dianoia Institute of Philosophy, A.C.U.

Causal models provide us with a formal tool for representing the networks of determination in which causes and effects are embedded. They tell us how some token features of the world—represented in the model with variables—determine others. They tell us whether one variable determines another along a single path or along multiple paths. They tell us whether two variables determine a third, and, if so, whether they do so along independent or intersecting paths. And it has been hoped that they can also tell us whether one variable value is a token cause of another.[1] To this end, a number of authors have developed theories of token causation in the causal modeling framework (e.g., Halpern and Pearl 2001, 2005;

For helpful conversations and feedback on this material, I am indebted to Gordon Belot, Daniel Drucker, Malcolm Forster, Jeremy Goodman, Christopher Hitchcock, James M. Joyce, Harvey Lederman, L. A. Paul, Brian Weatherson, Mark Wilson, and James Woodward—as well as audiences at the University of North Carolina, Chapel Hill; the Center for the Philosophy of Science at the University of Pittsburgh; and the Causal and Explanatory Reasoning conference at Venice International University. I am *especially* indebted to two anonymous reviewers for this journal, whose generous feedback and watchful eyes helped make this paper much better than it would otherwise have been.

1. Token causation is sometimes called 'singular causation' or 'actual causation'. Token causal relations are described by token causal claims—sentences of the form "*c*'s *F*-ing caused *e* to *G*" or "*c*'s *F*-ing was a cause of *e*'s *G*-ing," where *c*'s *F*-ing and *e*'s *G*-ing are token events (e.g., "Chris's drinking was a cause of his esophageal cancer"). These are to be contrasted with *type* or *general* causal claims like "Drinking causes esophageal cancer." So too should they be contrasted with the relations of causal determination between variables—the relations represented with directed edges in a causal graph. (Looking ahead to section 1, in figure 1, whether *B* fires causally determines whether *E* fires, but *B*'s failure to fire is not a token cause of *E*'s firing.) Throughout, "cause" should be understood to mean "token cause."

Philosophical Review, Vol. 130, No. 1, 2021
DOI 10.1215/00318108-8699682

Hitchcock 2001, 2007a; Woodward 2003; Menzies 2004, 2006; Hall 2007; Halpern 2008, 2016; Beckers and Vennekens 2017, 2018; Weslake, forthcoming; Andreas and Günther 2018, 2020). Lots of good work has been done on this front, but most of the theories developed to date have an awkward consequence: adding or removing an inessential variable from a model will lead these theories to revise their verdicts about whether two variable values are causally related or not.[2] Attend to an additional, inessential variable lying along a path from *C* to *E*, and these theories will change their mind about whether *C* caused *E*. Attend to an additional, inessential variable feeding into a path leading from *C* to *E*, and these theories will likewise change their minds about whether *C* caused *E*.[3]

I believe that this should concern us. In several instances, these theories are only able to agree with intuition through a judicious choice of which variables to include in the model. For just one example: in section 1.1 below, we'll encounter two systems which appear to differ causally, but which may be modeled with isomorphic variables and equations. Nonetheless, Christopher Hitchcock (2001) treats them differently by including an inessential variable in his model of one system while omitting the corresponding variable from his model of the other. There is a serious worry that, in the absence of some more general guidance about when variables can be ignored, and when not, ad hoc decisions can be used to effectively shield a theory from refutation. A theory whose causal verdicts don't change as inessential variables are added or removed—a *model-invariant* theory—would protect us from this kind of special pleading. Such a theory would have the added virtue of making it easier to determine whether *C* caused *E*. With such a theory, we needn't consider all possible correct causal models, nor decide which is most appropriate or apt for the present context; we need only check whether *C* caused *E* in a single correct model.

Below, I will provide a model-invariant theory of causation. Along the way, we'll see reason to think that an adequate theory of causation must distinguish between states which are normal, default, or inertial, and events which are abnormal, deviant departures therefrom

2. The theory of Sander Beckers and Joost Vennekens (2017, 2018) is a notable exception—modulo some finicky issues related to their 'timings'. Unfortunately, this theory says that a preemptive overdeterminer (see section 4) is not a cause. Beckers and Vennekens recognize and embrace this consequence of their theory, but it is not one that I am willing to endorse.

3. To understand why these theories are model-variant, see Gallow, n.d. I will get more precise about the term 'inessential' in section 2 below.

(section 1.1). This is striking even after you've been persuaded that it is true. Why should a distinction between default and deviant behavior play a role in our causal thought and talk? The theory developed here suggests an answer. In rough outline, the theory says that *C* caused *E* whenever both *C* and *E* are deviant or noninertial events, and there is an uninterrupted process which *transmits C*'s deviancy to *E*. That is, according to this theory, a cause is something which transmits aberrational behavior to its effect; and, if that is what a cause is, then it is no surprise to find the distinction between the default and the deviant, the normal and the abnormal, or the inertial and the noninertial showing up in our theorizing about causation.

In section 1, I will introduce causal models, show how they can be used to provide a semantics for causal counterfactual conditionals, and explain why I've been persuaded that these models must include information about which variable values are more default, normal, or inertial than which others. Then, in section 2, I will explain more carefully what I mean when I call a theory of causation formulated in terms of these causal models *model-invariant*. Sections 3–5 develop the notion of a *causal network*. This is a formal characterization of what I called above "an uninterrupted process which transmits *C*'s deviancy to *E*." Causal networks are the heart of my theory of causation, and they are model-invariant. In section 6.1, I will give some further motivation for thinking of causal networks as transmitting deviancy from cause to effect. In section 6.2, I will give a precise statement of the theory and illustrate it with some further applications.

A few words of forewarning: in what follows, I will for the most part confine my attention to some simple 'neuron systems' (see section 1 below)—though, along the way, I'll provide some 'real world' cases which exemplify similar causal structures. All of these systems will be deterministic. This narrow focus will allow me to sidestep some thorny questions—for instance, which kinds of variables can be included in a causal model, when a system of equations is correct,[4] and when one variable value is more or less default, normal, or inertial than another. By focusing on neuron systems, I will be able to get by with a small number of relatively uncontroversial assumptions about these contentious questions. Any complete theory of causation must say more about these issues than I will say here—just as it must be extended to cover indeterministic

4. I've said a bit about this in Gallow 2016, and I'll say a bit more in section 2 below, though there remains more to be said.

systems. Accordingly, the story I will tell here is an important part of a full theory of causation, but it is not yet a complete theory.

1. Causal Models

As I'll be using the term here,[5] a *causal model* consists of five components: a collection of exogenous variables,[6] **U**; an assignment of values to those variables, **u**; a collection of endogenous variables, **V**; a system of *structural equations*, one equation for each endogenous variable in **V**; and a specification of which of a variable's values are more *normal*, *typical*, *inertial*, or *default* than which others.[7]

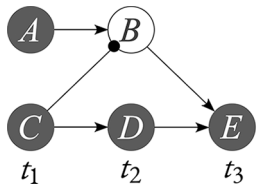
CAUSAL MODELS
A causal model $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$ is a 5-tuple of
(a) An m-tuple, $\mathbf{U} = (U_1, U_2, \ldots, U_m)$, of *exogenous* variables;
(b) An assignment of values, $\mathbf{u} = (u_1, u_2, \ldots, u_m)$, to **U**;
(c) An n-tuple, $\mathbf{V} = (V_1, V_2, \ldots, V_n)$, of *endogenous* variables;
(d) A system of *structural equations*, $\mathbf{E} = (\phi_1, \phi_2, \ldots, \phi_n)$, one equation for each endogenous variable $V_i \in \mathbf{V}$; and
(e) A specification, $\geq$, of which values of each variable in $\mathbf{U} \cup \mathbf{V}$ are more *default, normal, typical,* or *inertial* than which others.

To see how a causal model represents structures of causal determination, consider the Lewisian system of neurons shown in figure 1. Here's how to read the diagram in figure 1: for each time listed at the bottom, the neurons above it can either fire or not fire at that time. If a

5. This terminology is slightly idiosyncratic. Many authors do not include either $\geq$ or **u** in their definition of 'causal model'.

6. As I understand them, variables are functions from some domain to the real line—in my view, the domain is a set of possible spacetime regions. So, as I understand them, variables tell you what their possible values are (they are just the real numbers in the image of the function). The reader may think about variables differently; but they should ensure that a causal model tells us which values each variable may take on.

7. Notation: variables will be denoted with uppercase italic letters ($A, B, C, \ldots$), while their values will be denoted with the corresponding lowercase letters ($a, b, c, \ldots$). Tuples will be indicated with boldface. I will use uppercase for a tuple of variables and lowercase for a tuple of their values. The Greek letter ϕ, subscripted with a variable, will stand for a function, and I will often use just 'ϕ_V' to stand for an entire structural equation like $V := \phi_V(X, Y, Z)$. Throughout, I will apply set-theoretic notation to *tuples* of variables. Thus, $\mathbf{U} \cup \mathbf{V}$ is a tuple containing all and only the variables in either **U** or **V**, $\mathbf{V} \backslash \mathbf{X}$ is a tuple containing all and only the variables in **V**, except for those in **X**, and so on. There will in general be many such tuples, depending upon an arbitrary choice of order. It won't matter which of these an expression like '$\mathbf{U} \cup \mathbf{V}$' denotes. In sections 3–6, I will use calligraphic letters ($\mathcal{P}, \mathcal{N}$) to stand for sets of *directed edges*.



$$\mathbf{U}_1 : (A, C) \qquad \mathbf{u}_1 : (1, 1) \qquad \mathbf{V}_1 : (B, D, E)$$

$$\mathbf{E}_1 : \begin{pmatrix} E := B \vee D \\ D := C \\ B := A \wedge \neg C \end{pmatrix}$$

Figure 1. On the left, the neuron system *Preemptive Overdetermination*. On the right, the canonical causal model, $\mathbf{M}_1$, of this neuron system. (For all variables, 0 is default and 1 is deviant.)

neuron actually fires at its designated time, then it is colored gray. Otherwise, it is colored white. The arrows represent stimulatory connections between neurons. If the neuron at the tail of the arrow fires at its designated time, then, ceteris paribus, the neuron at the head will fire at its designated time. Thus, if either *B* or *D* in figure 1 fires at t_2, then *E* will fire at t_3. The circle-headed lines represent inhibitory connections between neurons. If the neurons at their base fire, then the neurons at their head definitely *won't* fire. In figure 1, for instance, if *C* fires at t_1, then *B won't* fire at t_2, no matter whether *A* fires or not.

Parenthetically, it is not uncommon to see diagrams like these used to represent the causal scenarios described in vignettes—scenarios involving rock throwings, coffee poisonings, and the like. This is not how I will be using them here. Rather, I will be understanding these diagrams as representing hypothetical mechanical systems obeying the simple laws described above. These systems consist of a small number of parts, the neurons, with two potential states: being *dormant*, which is a neuron's inertial state, the state in which it will remain unless acted upon from without, and *firing*, which a neuron will only do when another neuron connected to it with a stimulatory connection fires.[8] You could think of these diagrams as representing an appropriately connected electrical circuit (cf. Armstrong 2004: 446), neurons in the brain, connected with appropriate excitatory and inhibitory synapses,[9] or a boring possible world containing no more than a few objects, the 'neurons', and governed by simple laws of nature specifying when these neurons will and will not fire. I'll be using these neuron systems not as representational tools

8. Some neuron systems I introduce later on will have more potential states than these. I will explain the additional complications then.

9. This is how Lewis thought of them (e.g., Lewis 1986: 196).

but rather as the reality to be represented with a causal model (cf. Hitchcock 2007b).

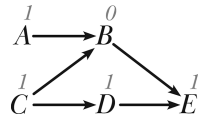
To represent the neuron system shown in figure 1, we may assign a variable to every neuron: *A, B, C, D,* and *E.* These variables take on the value 1 if their associated neurons fire at their designated times, and take on the value 0 if their associated neurons remain dormant at their designated times. (Thus, I use '*A*' for both the neuron and the variable which represents whether *A* fires at t_1. Context will disambiguate.) Both *A* and *C* are *exogenous* variables—variables whose values are not causally determined by the values of the other variables in the model. Since both of these neurons fire at t_1, the exogenous assignment will tell us that $A = C = 1$. The variables *B, D,* and *E* will be *endogenous*—variables whose values are causally determined by the values of other variables in the model. The structural equations in **E** tell us exactly *how* the values of the endogenous variables are causally determined. The equation $E := B \vee D$ tells us, firstly, that whether *E* fires is causally determined by whether *B* does and whether *D* does and, secondly, that *E* will fire if and only if (iff) either *B* or *D* does.[10] Similarly, the equation $D := C$ tells us that whether *D* fires is causally determined by whether *C* does, and that *D* will fire iff *C* does. The structural equations, together with the exogenous variable assignment, allow us to solve for the value of every variable in the model. For instance, in $\mathbf{M}_1$ (the model of the neuron system in figure 1), the structural equation $B := A \wedge \neg C$, together with the exogenous assignment $A = C = 1$, tells us that $B = 0$. Similarly, the structural equation $D := C$, together with the exogenous assignment $C = 1$, tells us that $D = 1$. And, finally, the structural equation $E := B \vee D$, together with the values $B = 0$ and $D = 1$, tells us that $E = 1$.

Because the equations in **E** encode information about the direction of causal determination, we cannot rearrange $D := C$ to get $C := D$, as we could with an ordinary equation. A *structural* equation $V := \phi_V(U)$ tells us more than just that the value of *V* is a function, ϕ_V, of the value of *U*. It additionally tells us that the value of *V* is *causally determined by* the value of *U*, in a way that the value of *U* is not causally determined by the value of *V*. This is why we use ':=', rather than '=', in structural equations.

Given a causal model, **M**, we may construct a *causal graph* which displays the causal determination structure among the variables in the model, as follows: if a variable *U* appears on the right-hand side of a

10. Notation: $X \wedge Y$, $X \vee Y$, and $\neg X$ are the Boolean functions $\min\{X, Y\}$, $\max\{X, Y\}$, and $1 - X$, respectively.

variable V's structural equation, ϕ_V, then place a directed edge between U and V, with its tail at U and its head at V, $U \rightarrow V$. Thus, given the causal model shown in figure 1, we may construct the following causal graph:



(Note: I have additionally decorated the graph with the values the variables take on in the model.) This graph tells us that the variables A and C are exogenous, that B's value is causally determined by the values of A and C, that D's value is causally determined by the value of C, and that E's value is causally determined by the values of B and D. While it tells us *by which* other variables each endogenous variable is causally determined, the graph on its own does not tell us *how* the values of the endogenous variables are causally determined. For that information, we must look to the structural equations in **E**.

It is common to use the metaphor of genealogy to describe the causal determination relations between variables displayed in a graph. For instance, B and D are E's causal parents, and C's causal children. Similarly, B, D, and E are C's causal descendants. Throughout, I will assume that no variable is among its own causal descendants—that is, I will assume that there are no causal loops.[11] I will use '**PA**(V)' to denote a tuple of V's causal parents.

Finally, our causal model should specify, for each variable, which values of that variable are more *default*, *inertial*, or *normal* than which others. In the case of the neuron system from figure 1, I will assume that remaining dormant is the default, normal state of a neuron—it is the state in which the neuron will remain unless it is acted upon by some other, stimulatory neuron. And I will assume that firing is a more abnormal deviation from that default, inertial state. I will assume likewise for every other neuron system in this paper.[12] The reader may be curious why

11. I make this assumption in the interests of simplicity, not out of necessity. Local dependence (see section 4) is well defined in cyclic models; so causal networks (see section 5) are well defined in cyclic models; so the theory of causation I'll present in section 6 can be applied straightforwardly to cyclic models.

12. Formally, we can understand $\geq$ as a function from the variables $V \in \mathbf{U} \cup \mathbf{V}$ to a partial pre-order over their values, $\geq_V$. If $v \geq_V v^*$, then v is no more default, normal, or inertial than v^* (cf. Halpern 2008, 2016; Halpern and Hitchcock 2015). Perhaps which

this kind of information is included in a causal model; I'll explain in section 1.1 below.

1.1. Defaults and Deviancy

The neuron system shown in figure 1 gives a case of *Preemptive Overdetermination*. There, either *A*'s firing or *C*'s firing would have been enough, on its own, to make *E* fire. Both *A* and *C* fired, so the firing of *E* was overdetermined. But the overdetermination is not symmetric. Though the causal process initiated with *C*'s firing runs to completion, the causal process initiated with *A*'s firing is *preempted* by *C*'s firing. *A* would have caused *E* to fire, were it not for *C*; but, as it happens, *A* is merely a backup, would-be cause. *C*, on the other hand, is a genuine cause of *E*'s firing.

Consider the neuron system shown in figure 2. (I follow Ned Hall [2007] in calling this neuron system a 'short circuit'.)[13] There, the neuron *C* fires, causing *B* to fire; and *B*'s firing threatens to make *E* fire. But, at the same time that *C* initiates this threat to *E*'s dormancy, it also makes *D* fire. And *D*'s firing prevents *E* from firing. So *C* both creates a threat to *E*'s dormancy and, at the same time, neutralizes that very threat. For a case with a similar causal structure, consider:[14]

> *Boulder*
> Matthew hikes through the Scottish highlands. Above him, a large boulder becomes dislodged and careens down the hillside. He sees the boulder coming and jumps out of the way at the last second, narrowly escaping death.

variable values are more inertial than which others should be relativized to the values of some other variables in the model. Taking for granted that your food is poisoned, your death may be inertial, even though, when we don't take this for granted, death is an abnormal departure from inertial behavior (cf. Halpern 2016). Perhaps we should further distinguish variable values which are *inertial* from those which are *deviant*, saying that, conditional on the poisoning, your death is inertial, but deviant. I'm sympathetic to these thoughts; but I'll put them aside for the nonce. We will be able to say many interesting things without worrying too much about the particulars of the default/deviant distinction.

13. See also Lewis 2004: 97–99, in which the same structure is called an *inert network*.

14. This case is attributed to an early draft of Hall 2004 by Hitchcock (2001). In assuming that *Boulder* and *Short Circuit* have similar causal structures, I am in part assuming that the boulder's fall is a deviant, noninertial event, and that Matthew's surviving is a default, inertial state.

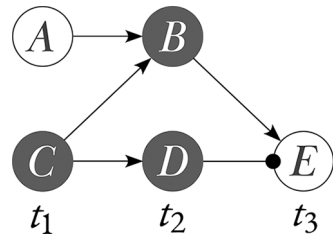
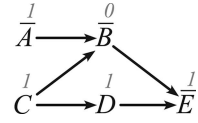


Figure 2. *Short Circuit.*

The boulder's becoming dislodged creates a threat to Matthew's life. However, at the same time that it creates this threat, it also alerts him to its presence, causing him to jump out of the way. So the boulder both creates a threat to Matthew's life and, at the same time, neutralizes that very threat. I take it that the boulder's becoming dislodged did not cause Matthew to survive—and I take it that *C*'s firing did not cause *E* to remain dormant in the neuron system from figure 2.

As Hall (2007) notes, we may write down a system of structural equations for *Short Circuit* which is isomorphic to the model of *Preemptive Overdetermination* from figure 1. Let $\overline{A}$ be a variable which takes on the value 1 if the neuron *A doesn't* fire, and takes on the value 0 if it *does*. Similarly, let $\overline{B}$ and $\overline{E}$ be variables which take the value 1 if their associated neurons *don't* fire, and take on the value 0 if they *do*. And let *C* and *D* be variables which take on the value 1 if their associated neurons fire and take on the value 0 if they don't. Then, the following system of equations will correctly describe the causal determination structure among these variables.

$$\overline{E} := \overline{B} \vee D$$
$$D := C$$
$$\overline{B} := \overline{A} \wedge \neg C$$



E won't fire just in case either *B* doesn't fire or *D* does; *D* will fire just in case *C* does; and *B won't* fire just in case neither *A* nor *C* do.

These are isomorphic to the equations we wrote down for the case of *Preemptive Overdetermination*. Moreover, the exogenous variables take on precisely the same values. In *Preemptive Overdetermination*, *C*'s firing caused *E* to fire (that is, $C = 1$ caused $E = 1$). But, in *Short Circuit*, *C*'s firing did not cause *E* to not fire (that is, $C = 1$ did not cause $\overline{E} = 1$). So, if we wish to use causal models to determine which variable values caused which other

variable values, then we will need to know more than a true system of structural equations and an assignment of values to the exogenous variables is capable of telling us.

It is natural to think of the dormancy of a neuron as a kind of default, normal, or inertial state. It is the state in which the neuron will remain unless it is acted upon by some other, stimulatory neuron. And the event of a neuron's firing is a deviation from that default, normal, inertial state. Several authors have thought that this distinction, between *default*, *normal*, or *inertial* states and events which are *abnormal*, *noninertial deviations* therefrom, must be incorporated into a theory of causation.[15] And appealing to this distinction allows us to distinguish *Preemptive Overdetermination* from *Short Circuit*. For, in our model of *Preemptive Overdetermination*, $A = 1$, $B = 1$, and $E = 1$ stand for the *deviant*, *abnormal*, *noninertial* events of neurons firing; while, in our model of *Short Circuit*, $\overline{A} = 1$, $\overline{B} = 1$, and $\overline{E} = 1$ stand for the *default*, *normal*, *inertial* states of neurons remaining dormant. It is for this reason that a causal model includes $\geq$, which tells us which variable values are more deviant, abnormal, or noninertial than which others.

No theory of causation incorporating this kind of information is complete until it provides an independent characterization of which variable values are more or less default than which others.[16] However, insofar as we keep our focus on simple neuron systems, the only assumption I will need is that a neuron's remaining dormant is more default than its firing. When additional assumptions about the deviancy of a variable's values are needed, I will explicitly state them.

The focus on simple neuron systems will also allow me to get by with just one, relatively weak, assumption about when a causal model is correct. To understand this assumption, return to the neuron system shown in figure 1. To construct the causal model $\mathbf{M}_1$ from this neuron system, we assigned a variable to every neuron, with a value of 1 standing for the neuron firing at its designated time, and a value of 0 standing for the neuron remaining dormant at that time. The variables for the neurons on the far left-hand side—those without any stimulatory or inhibitory connections coming into them—were made exogenous, and assigned

15. See in particular Kahneman and Miller 1986; Thomson 2003; Maudlin 2004; McGrath 2005; Hall 2007; Hitchcock 2007a; Halpern 2008, 2016; Hitchcock and Knobe 2009; Paul and Hall 2013; Halpern and Hitchcock 2015.

16. For some attempts, see Kahneman and Miller 1986; Maudlin 2004; McGrath 2005; Hall 2007; Hitchcock 2007a; Hitchcock and Knobe 2009; Wolff 2016.

the values corresponding to the actual state of their associated neurons. We then wrote down equations describing how the state of each endogenous neuron is directly causally determined by the other neurons in the system, and we assumed that firing is a more deviant state of a neuron than remaining dormant. Let's call the causal model that we construct in this way from a given neuron system the *canonical model* of that neuron system. Then, the assumption I'll need about model correctness going forward is this: the canonical causal model of any neuron system is correct.

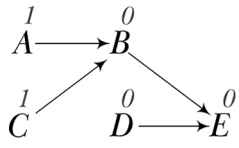
1.2. Counterfactual Causal Models

Given a causal model $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, with some tuple of variables $\mathbf{A} \subseteq \mathbf{U} \cup \mathbf{V}$, we may construct a *counterfactual* model, in which the variables in **A** have been intervened upon to set their values to **a**, as follows: we *remove* any endogenous variables in **A** from the endogenous variables, **V**, and add them to the exogenous variables, **U**. Next, we remove the structural equations of any endogenous variables in **A** from the system of structural equations, **E**, and change the exogenous assignment **u** so that it assigns the values in **a** to the variables in **A**. The information in $\geq$ will remain unchanged.

> Counterfactual Causal Model
> Given a causal model $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, including the variables in **A**, and given the assignment of values **a** to **A**, the counterfactual model
> $$\mathbf{M}[\mathbf{A} \rightarrow \mathbf{a}] = (\mathbf{U}^*, \mathbf{u}^*, \mathbf{V}^*, \mathbf{E}^*, \geq^*)$$
> is the model such that:
> (a) $\mathbf{U}^* = \mathbf{U} \cup \mathbf{A}$
> (b) $\mathbf{u}^* = \mathbf{u} + \mathbf{a}$[17]
> (c) $\mathbf{V}^* = \mathbf{V} \backslash \mathbf{A}$
> (d) $\mathbf{E}^* = \mathbf{E} \backslash (\phi_A \mid A \in \mathbf{A})$
> (e) $\geq^* = \geq$

For instance, figure 3 displays the counterfactual model $\mathbf{M}_1[D \rightarrow 0]$, in which we have intervened to set *D*'s value to 0. Notice that, in this model, it is no longer the case that *D*'s value is causally determined by *C*'s. Rather, *D* has been 'exogenized', and it has been given the exogenous assignment

17. Here, I use '**u** + **a**' to refer to the result of adding the assignment from **a** to **u** (if the variable from **A** is not already exogenous) or revising the assignment **u** to match **a** (if the variable from **A** is already exogenous).



$$\mathbf{U}_1^* : (A, C, D) \qquad \mathbf{E}_1^* : \begin{pmatrix} E := B \vee D \\ B := A \wedge \neg C \end{pmatrix}$$
$$\mathbf{u}_1^* : (1, 1, 0)$$
$$\mathbf{V}_1^* : (B, E)$$

Figure 3. On the right, the counterfactual model $\mathbf{M}_1[D \rightarrow 0]$ (for all variables, 0 is default and 1 is deviant). On the left, its associated causal graph.

0. In this new model, when we solve for the values of the variables as before, we find that $E = 0$.

We can use these counterfactual models to provide a semantics for causal counterfactual conditionals.[18] According to this semantics, a causal counterfactual "Had **A** taken on the values **a**, then it would have been that C" (where C is any Boolean function of variable values) is true in a causal model **M** just in case C is true in the *counterfactual model* in which you've intervened to set the variables in **A** to the values **a**, $\mathbf{M}[\mathbf{A} \rightarrow \mathbf{a}]$.

> Causal Counterfactuals
> If C is a proposition about the values of the variables in a causal model **M**, and **M** contains the variables in **A**, then the causal counterfactual $\mathbf{A} = \mathbf{a} \; \Box\!\!\rightarrow$ C is true in **M** iff C is true in the counterfactual model $\mathbf{M}[\mathbf{A} \rightarrow \mathbf{a}]$,[19]

$$\mathbf{M} \models \mathbf{A} = \mathbf{a} \; \Box\!\!\rightarrow \mathrm{C} \Leftrightarrow \mathbf{M}[\mathbf{A} \rightarrow \mathbf{a}] \models \mathrm{C}$$

Thus, because $E = 0$ is true in the counterfactual model $\mathbf{M}_1[D \rightarrow 0]$, the counterfactual $D = 0 \; \Box\!\!\rightarrow E = 0$ is true in the model $\mathbf{M}_1$.

2. Model Invariance

Like any other vehicle of representation, a causal model may be appraised for accuracy. The model tells us that the world is a certain way, and what it tells us could be either true or false. In the former case, the model is correct. In the latter case, it is incorrect. A causal model which says that the rain is causally determined by the state of my umbrella is not correct; it gets the causal structure of the world backwards. Among the correct causal models, some are more detailed, some less so. One correct model tells us that whether the match lights is causally determined by whether it is struck. Another tells us that whether the match lights is causally deter-

18. For more on this semantics, see Galles and Pearl 1998; Briggs 2012; Huber 2013.

19. I use "$\mathbf{M} \models$ S" for "the sentence S is true in the model **M**." For sentences of the form "$V = v$," and Boolean functions of these sentences, the definition of truth in a model is just what you would expect.

mined both by whether it is struck and whether there is oxygen present. Both models tell us true things about the world's causal structure, though the second tells us strictly more. Other correct causal models may tell us which variables causally determine whether the match is struck, which are causally determined by whether the match is lit, and which are causally intermediate between the match's striking and its lighting.

If we wish to theorize about causation in terms of causal models, then it is important for us to distinguish between correct and incorrect models; for it is only the verdicts issued about *correct* models which are commitments of our theory. Without some way of distinguishing correct models from incorrect models, a theory of causation would tell us nothing at all about which variable values are token causes of which others.

For my purposes, I won't need to supply a complete account of when a causal model is correct. I will only need to endorse three, rather weak, conditions on the correctness of a causal model (namely, that the canonical causal model of a neuron system is correct, and the principles **Exogenous** and **Endogenous Removal**, to be introduced below). However, just to orient the reader, let me say a few things here about what I think it takes for a causal model to represent the world correctly.

On my view, in order to be correct, a causal model must entail only true counterfactuals about the values of the variables appearing in the model. If a causal model entails a false counterfactual, then the model is incorrect. But entailing only true counterfactuals is not sufficient for a model being correct; some incorrect models entail only true counterfactuals. To appreciate this, return to the case of *Preemptive Overdetermination* from figure 1, and consider a model which contains only the variables C and E, both of which are exogenous and take on the value 1. This model tells us, truly, that E's firing is counterfactually independent of C's firing, and that C's firing is counterfactually independent of E's. But it also tells us, falsely, that whether E fires is *causally* independent of whether C does. So this model is not correct, even though it entails only true counterfactuals. Or consider a model of *Preemptive Overdetermination* which contains only the variables A, C, and E, where A and C are exogenous and both take on the value 1, and a single structural equation which tells us that E fires iff either A or C does: $E := A \vee C$. This model will entail only true counterfactuals. However, in this model, the variables A and C are perfectly symmetric. So any theory of causation presented with this model will tell us that $A = 1$ caused $E = 1$ iff $C = 1$ caused $E = 1$. Since $C = 1$ caused $E = 1$ and $A = 1$ did not, this model cannot be correct. My diagnosis is that this too-simple model tells us, falsely, that A and C determine the value of E

along nonintersecting paths. So, on my view, causal models don't just represent patterns of counterfactual dependence between variable values—they also tell us something about the paths by which variables causally determine the values of their descendants.[20]

In general, my view is that a causal model tells us how each of the values of each endogenous variable, $V \in \mathbf{V}$, is causally determined by the values of V's ancestors in the model—in particular, whether they are determined by a single path or multiple paths, and whether they are determined by independent or intersecting paths—and it tells us that those values are *not* causally determined by V's nonancestors. From these facts, we can determine which variable values counterfactually depend upon which others, as described in section 1.2. So, if a model entails false counterfactuals, then the model must have told us something false. But the model tells us strictly more than these counterfactuals do. (I will expand upon this view when discussing some examples below.)

2.1. Exogenous Removal

In order to be correct, a causal model needn't include a variable for every factor which is potentially causally relevant. The model which says that whether the match lights is causally determined by whether it is struck and whether oxygen is present is correct. But, so long as oxygen *is* present, the variable for oxygen is not needed. We could remove it, and the causal model left behind—the one which tells us that whether the match lights is causally determined by whether it is struck—would be correct, also. (This model no longer tells us whether there's oxygen present, nor whether the presence of oxygen causally determines the lighting of the match, but no model will tell us *everything* about the world's causal structure, just as no map will tell us everything about where things are located. A map of London is not incorrect simply because it doesn't tell us where Sabeen's flat and the Eiffel tower are located. Likewise, a causal model is not incorrect simply because it doesn't tell us something about the values of omitted variables.) Or consider the neuron system displayed in figure 4. The canonical model of this neuron system, $\mathbf{M}_4$, includes a variable for A, C, and E (with 1 corresponding to firing and 0 corresponding

20. Again, this is my diagnosis of why the model is not correct; but if the reader disagrees with it, this disagreement won't make a difference to anything else I have to say here. My goal in this section is just to defend the principles **Exogenous** and **Endogenous Removal** (see below). And you can accept these principles while disagreeing with me about why this simple model of *Preemptive Overdetermination* is incorrect.

to not firing). Its exogenous assignment tells us that $A = 1$ and $C = 0$, and it includes the structural equation $E := A \wedge \neg C$. The canonical model $\mathbf{M}_4$ is correct; but, since C doesn't fire, the variable for C is not necessary. Just as we can take the presence of oxygen for granted, so too can we take the nonfiring of C for granted. So we can pluck the variable C out of the model and replace it with its actual value, 0, in the structural equation. We will be left with a model—call it '$\mathbf{M}_4^{-C}$'—which contains the sole exogenous variable A, the sole endogenous variable E, and the structural equation $E := A \wedge \neg 0$, or just $E := A$.

In general, if $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$ is a causal model with the exogenous variable $U \in \mathbf{U}$, then let $\mathbf{M}^{-U}$ be the model that you get by (a) removing U from $\mathbf{U}$; (b) removing U's value from $\mathbf{u}$; (c) 'exogenizing' any variables in $\mathbf{V}$ whose only parent was U;[21] (d) replacing U with its value in every structural equation in $\mathbf{E}$; and (e) removing information about the deviancy of U's values from $\geq$.

In my view, removing an exogenous variable from a correct causal model in this way will not always leave a correct causal model behind. For instance, consider the neuron system in figure 5. This is just like the neuron system from figure 4, except that, in figure 5, C fires, and therefore, E doesn't. The canonical model of this neuron system, $\mathbf{M}_5$, will be exactly like $\mathbf{M}_4$, except that the exogenous assignment will tell us that $C = 1$, rather than $C = 0$. In my view, this makes a difference with respect to whether the variable C can be ignored. For if we try to replace C with its actual value in $\mathbf{M}_5$, we will be left with the structural equation $E := A \wedge \neg 1$, which is a constant function of A. Whether A is 0 or 1, E will take on the value 0. This equation tells us, falsely, that E and A are causally independent. So the model $\mathbf{M}_5^{-C}$ is not correct, even though $\mathbf{M}_5$ is. So removing an exogenous variable does not always preserve correctness.

In my view, in order for a structural equation $V := \phi_V(\mathbf{PA}(V))$ to be correct, it must tell us how each of V's values are causally determined by the values of V's causal parents. So ϕ_V must be a *surjective* function of *all* of the right-hand-side variables. That is: for every value v of the left-hand-side variable V, there must be some assignment of values to the right-hand-side variables $\mathbf{PA}(V)$ which gets mapped to v by the function ϕ_V. If ϕ_V is not surjective, then the structural equation for V cannot tell us how each of V's values could be causally determined by the values of V's

21. 'Exogenizing' a variable $V \in \mathbf{V}$ means (a) moving V from $\mathbf{V}$ to $\mathbf{U}$; (b) enriching the exogenous assignment $\mathbf{u}$ so that it assigns V the value it takes on in the original model $\mathbf{M}$; and (c) removing V's structural equation from $\mathbf{E}$.

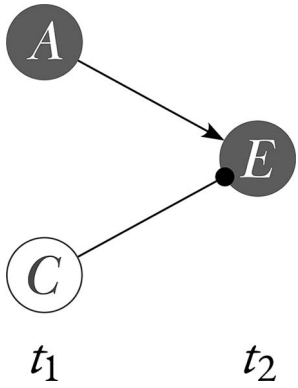


Figure 4. *Omission.*

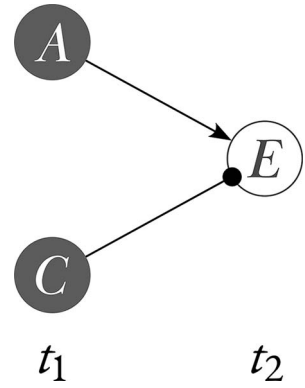


Figure 5. *Prevention.*

parents. So, if ϕ_V is not surjective, then the structural equation for V cannot be correct.[22] Additionally: on my view, a structural equation ϕ_V tells us that the left-hand-side variable has its value causally determined by *all* of the right-hand-side variables. So ϕ_V must be a *function* of all of V's causal parents. That is: for each $Pa \in \mathbf{PA}(V)$, there must be some assignment of values to the *other* variables in $\mathbf{PA}(V)$ such that, when the other parents take on those values, the value V takes on depends upon which value Pa takes on.

In general, if U is exogenous in $\mathbf{M}$, and if every structural equation ϕ_V in $\mathbf{M}^{-U}$ is both (a) a surjective function and (b) a function of all of V's remaining causal parents, then I will say that U is an *inessential* exogenous variable in $\mathbf{M}$.[23] Though removing exogenous variables will not always preserve correctness, I believe that removing *inessential* exogenous variables will. That is, I believe we should endorse the following principle.[24]

22. Or so it seems to me. You may not agree that structural equations must be surjective. If so, this shouldn't prevent you from accepting anything else I have to say here. By imposing this requirement, I strengthen the antecedent of my principle **Exogenous Removal** (see below). Strengthening the antecedent weakens the conditional. I think that this weakening is necessary; but if you think the principle is weaker than it needs to be, this is no reason for you to worry about its truth. (Readers who worry about this surjectivity requirement should also note that it is not required at any point in the proof of my theory's model-invariance in the appendix. So the theory would still be model-invariant even if **Exogenous Removal** were strengthened. Indeed, the theory would be model-invariant even if we say that *every* exogenous variable is inessential.)

23. V's *remaining* causal parents in $\mathbf{M}^{-U}$ are just V's causal parents in $\mathbf{M}$, minus U.

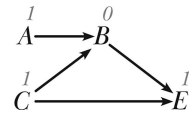
24. To be clear: I think that **Exogenous Removal** is a substantive claim; you could very well disagree with me about it (if, for instance, you thought that removing any exogenous

Exogenous Removal
If a causal model $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, is correct, and $U \in \mathbf{U}$ is inessential, then $\mathbf{M}^{-U}$ is also correct.

2.2. *Endogenous Removal*

In order to be correct, a causal model need not include a variable for every factor which is causally intermediate between two variables. Whether the room is illuminated is causally determined by whether the switch is up. There are ever so many variables causally intermediate between these two—whether current is flowing, whether the filament in the bulb is heated, and so on. Nevertheless, a model which omits them all is still correct. So, just as we may remove inessential exogenous variables from a causal model, so too may we remove inessential endogenous variables. Consider again the model $\mathbf{M}_1$, shown in figure 1. This model tells us that whether E fires is determined by whether D does, and that whether D does is determined by whether C does. Here, the variable for D is not necessary. We could pluck it out of the model by replacing it with the right-hand side of its structural equation, C, wherever it appears. We will be left with a model—call it '$\mathbf{M}_1^{-D}$,'—which contains the following system of structural equations.

$$E := B \vee C$$
$$B := A \wedge \neg C$$



This model won't tell us how D fits into the causal determination structure of the neuron system, but it tells us about the causal determination structure among the variables A, B, C, and E, and what it tells us about them is all correct.

In general, if $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$ is a causal model with the endogenous variable $V \in \mathbf{V}$, then let $\mathbf{M}^{-V}$ be the model that you get by (a) leaving $\mathbf{U}$ and $\mathbf{u}$ alone; (b) removing V from $\mathbf{V}$; (c) removing V's structural equation $V := \phi_V(\mathbf{PA}(V))$ from $\mathbf{E}$; (d) replacing V with $\phi_V(\mathbf{PA}(V))$ wherever V appears on the right-hand side of a structural equation in $\mathbf{E}$; and (e) removing information about V from $\geq$.

variable with a deviant value wouldn't leave a correct model behind). I don't intend for **Exogenous Removal** to be an implicit partial definition of what I mean by 'correctness'.

In my view, removing an endogenous variable from a correct causal model in this way will not always leave a correct causal model behind. As with exogenous variables, removing some endogenous variables won't leave behind surjective, functional structural equations. These variables are not inessential. But they are not the only ones. Consider again the model $\mathbf{M}_1^{-D}$. If we pluck the variable B out of this model in the manner specified above, then we will arrive at a model, $\mathbf{M}_1^{-D,-B}$, which contains the sole structural equation $E := (A \wedge \neg C) \vee C$, or just $E := A \vee C$, and the exogenous assignment $A = C = 1$. This model treats the variables A and C symmetrically; yet A and C differ causally. So the model $\mathbf{M}_1^{-D,-B}$ cannot be correct. As I remarked above, in my view, this is because $\mathbf{M}_1^{-D,-B}$ tells us that A and C causally determine the value of E along nonintersecting paths, which is not true—but, whatever the reason, we should agree that the model is incorrect, since $C = 1$ caused $E = 1$ and $A = 1$ did not.

Suppose that, in $\mathbf{M}$, V has a single parent, *Pa*, and a single child, *Ch*, $Pa \rightarrow V \rightarrow Ch$, and suppose that *Pa* is not *also* a parent of *Ch*. If that's so, then say that V is an *interpolated* variable in $\mathbf{M}$. If V is interpolated, then I'll say that it is an inessential endogenous variable.[25] Though removing endogenous variables will not always preserve the correctness of a causal model, I believe that removing inessential endogenous variables will. That is, I think we should endorse the following principle.

> **Endogenous Removal**
> If a causal model $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$ is correct, and $V \in \mathbf{V}$ is inessential, then $\mathbf{M}^{-V}$ is also correct.

2.3. *Model-Invariance*

We want a theory which will tell us whether two variable values, $C = c$ and $E = e$, are causally related, and we wish to formulate that theory within the framework of causal models. (Throughout, I use 'C' and 'E' for the cause and effect variables of interest, and 'c' and 'e' for their actual values.) This theory will say whether $C = c$ caused $E = e$ *relative to* a given causal model. For an arbitrary C and E, there will be a great many correct causal models containing both C and E. It would be nice if our theory did not require us to survey them all. It would be nice if its verdicts did not vary from correct

25. Note that, if V is interpolated, then all of the equations in $\mathbf{M}^{-V}$ will automatically be surjective functions of all of their right-hand-side variables, so long as all of the equations in $\mathbf{M}$ are.

model to correct model. That is, it would be nice if our theory satisfied the following constraint.[26]

> **Model Invariance**
> For any two causal models **M** and **M′** which both contain the variables C and E, if both **M** and **M′** are correct, then $C = c$ caused $E = e$ in **M** iff $C = c$ caused $E = e$ in **M′**.

Let's call a theory of causation which is consistent with the principles **Model Invariance**, **Exogenous Removal**, and **Endogenous Removal** a *model-invariant* theory of causation.[27] If a theory is inconsistent with these principles, then let's say that it is a *model-variant* theory of causation. It would be nice to have a model-invariant theory. If our theory is model-invariant, then, when we ask whether $C = c$ caused $E = e$, we needn't worry about our verdict changing as we include additional variables lying along, or feeding into, paths from C to E. Nor need we worry about the theory being shielded from refutation by ad hoc choices about which variables to include and which to ignore. Unfortunately, almost all of the extant theories of causation in the causal modeling

26. There are alternatives to accepting **Model Invariance**. In general, let us say that a theory of causation formulated with causal models specifies when a causal model is a *witness* to $C = c$ causing $E = e$. We might go on to say that $C = c$ caused $E = e$ iff there is *some* witness to $C = c$ caused $E = e$ (and therefore, $C = c$ didn't cause $E = e$ iff there is *no* witness). Or we might say that $C = c$ caused $E = e$ iff *all* correct models containing C and E are witnesses to $C = c$ causing $E = e$ (and therefore, $C = c$ didn't cause $E = e$ iff *some* correct model fails to witness $C = c$ causing $E = e$). The first alternative makes it easy to establish causation but difficult to establish noncausation (we must establish noncausation in all of the correct models). Likewise, the second alternative makes it easy to establish noncausation, but difficult to establish causation. Model-invariance makes it easy to establish causation and noncausation both.

Cf. Joseph Y. Halpern (2016: sec. 4.4), who shows that his theory of causation will not revise its judgments of *non*causation as endogenous variables are removed, though it may reverse its judgments of causation. (Note that this result requires strong assumptions about normality. Given the assumption that, ceteris paribus, a neuron's firing is more deviant than its remaining dormant, Halpern's theory will reverse its verdicts about noncausation as well. See Gallow, n.d.)

27. Note that a theory's verdicts about causation will be preserved when inessential variables are removed iff that theory's verdicts about *noncausation* are preserved when inessential variables are *added*. And a theory's verdicts about noncausation will be preserved when inessential variables are removed iff that theory's verdicts about *causation* are preserved when inessential variables are *added*. So, if we are able to show that a theory's verdicts about both causation and noncausation don't change when inessential variables are *removed*, we will have also thereby shown that its causal verdicts don't change when inessential variables are *added*.

framework are model-variant. In particular, the accounts of Hitchcock (2001, 2007), Joseph Y. Halpern and Judea Pearl (2001, 2005), James Woodward (2003), Halpern (2008, 2016), Brad Weslake (forthcoming), and Holger Andreas and Mario Günther (2018, 2020) will all reverse or suspend their verdicts when inessential variables are added to or removed from a causal model (see Gallow, n.d.).

In sections 3–6, I will introduce a theory of causation that is model-invariant. If this theory says that $C = c$ caused $E = e$ in a causal model **M**, then it will continue to say this after any inessential variables are removed from **M**. And, if this theory says that $C = c$ *didn't* cause $E = e$ in **M**, then it will continue to say this after any inessential variables are removed from **M**. I will introduce this theory by walking through some standard problem cases from the literature—symmetric overdetermination (in section 3), preemptive overdetermination (in section 4), and counterexamples to transitivity (in section 5). According to this theory, a cause must be connected to its effect by what I will call a 'causal network'. In rough outline, a causal network represents an uninterrupted process, each stage of which depends upon its predecessors, and which *transmits* the cause's deviant, noninertial behavior to the effect. The definition of a causal network will be developed in section 5. Then, in section 6, I will state the theory, apply it to some additional cases, and try to motivate thinking of a causal network as a process which transmits deviant, noninertial behavior.

3. Symmetric Overdetermination

A simple case of symmetric overdetermination is shown in figure 6. Either *A*'s or *C*'s firing would have been enough, on its own, to make *E* fire. Both *A* and *C* fired, so the firing of *E* was overdetermined, and symmetrically so. There's nothing that *A*'s firing has that *C*'s firing lacks; nor anything *C* has that *A* lacks. If either of them caused *E* to fire, then both of them did. For another case with a similar structure, consider *Pay Raise*.[28]

> *Pay Raise*
> Franny, Sammy, and Tammy vote on a proposal to raise legislators' salaries.

28. Cf. Livengood 2013. Note: when I say that *Pay Raise* has a similar causal structure to *Symmetric Overdetermination*, I am in part assuming that the 'yea' votes and the proposal's passing are deviant. (Of course, the causal structures are *similar*, not exactly the same. In *Symmetric Overdetermination*, *E* would still have fired, even if either *A* or *C* had not fired; and, in *Pay Raise*, the proposal would still have passed, even if either Franny, Sammy, or Tammy had not voted 'yea'.)

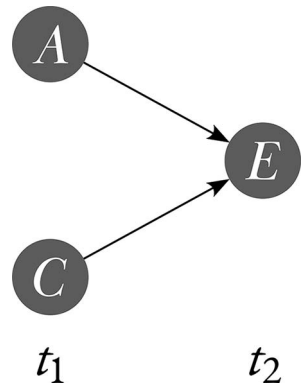


Figure 6. *Symmetric Overdetermination.*

> The proposal requires two out of three votes in order to pass. All three vote in favor, and the proposal passes.

The passing of the proposal was overdetermined by the three votes in favor, and symmetrically so. There's nothing that any one vote has that the others lack. If any vote caused the motion to pass, then all of them did.

In cases like these, the effect is *overdetermined*. The world supplied more than enough for the effect to obtain. There is some appeal to the idea that the world did this by supplying more than enough *causes*—that is, there is some appeal to the idea that each of the overdeterminers is individually a cause of the effect. For instance: *C* individually caused *E* to fire, and Franny individually caused the proposal to pass. At the same time, there is some appeal to the idea that *C*'s firing didn't *all by itself* cause *E* to fire, and that Franny didn't *all by herself* cause the proposal to pass. Perhaps she is *part* of a cause—perhaps she *contributed* to the proposal's passing—but, we may think, she did not cause it to pass all by herself, given that the proposal would have had a two-vote majority even without her support.

John L. Mackie (1965) and David K. Lewis (1986) were both happy with the verdict that *C*'s firing didn't cause *E* to fire in figure 6. According to both, in cases of symmetric overdetermination, intuition is split and a theory of causation could reasonably answer with either verdict.[29] I agree with Mackie and Lewis.[30] An adequate theory of causation needn't say

29. "Our ordinary concept of cause does not deal clearly with cases of this sort" (Mackie 1965: 251). "Such cases can be left as spoils to the victor, in D. M. Armstrong's

that C's firing caused E to fire. However, it should not say that E's firing was uncaused. If neither A nor C individually caused E to fire, then they must have done so *jointly*. I will formally represent A and C's jointly causing E by allowing not just individual variable values, but also *tuples* of variable values, to be causes. In the canonical model $\mathbf{M}_6$, to say that A's firing and C's firing jointly caused E to fire is to say that $(A, C) = (1, 1)$ caused $E = 1$.[31]

My theory will not say that C's firing individually caused E to fire. So I will take the lesson of *Symmetric Overdetermination* to be this: we should want a theory of causation to tell us more than whether an *individual* variable value $C = c$ caused a variable value $E = e$. We should also want it to tell us when some *collection* of variable values, $\mathbf{C} = \mathbf{c}$, caused a variable value $E = e$. That is: we should want a theory not just of *individual* causation, but of *joint* causation as well.[32]

Throughout, by the way, I will draw no distinction between a variable, V, and a 1-tuple containing that variable, (V)—nor will I distinguish between a variable value $V = v$ and a 1-tuple variable value $(V) = (v)$. This conflation allows a theory of joint causation to cover individual causation as a special case.

Once we allow tuples of variables to be causes, we should generalize **Model Invariance**. So generalized, the principle will tell us that, if both $\mathbf{M}$ and $\mathbf{M}'$ are correct and contain the variables in $\mathbf{C} \cup (E)$, then $\mathbf{C} = \mathbf{c}$ caused $E = e$ in $\mathbf{M}$ iff $\mathbf{C} = \mathbf{c}$ caused $E = e$ in $\mathbf{M}'$. This is how I will understand the principle, and the corresponding property of *model-invariance*, from here on out.

phrase. We can reasonably accept as true whatever answer comes from the analysis that does best on the clearer cases" (Lewis 1986: 194).

30. This view is increasingly unpopular. Halpern and Pearl, Hitchcock, Woodward, and Weslake, among others, take it as a desideratum of a theory of causation that it says that C's firing caused E to fire in figure 6. See also the arguments for this conclusion in Schaffer 2003.

31. '(A, C)' is a pair whose first component is the variable A and whose second component is the variable C. '$(1, 1)$' is a pair whose first and second components are both the value 1. Thus, '$(A, C) = (1, 1)$' says that $A = 1$ and $C = 1$.

32. We could try to generalize further by asking when one tuple of variable values, $\mathbf{C} = \mathbf{c}$, caused another, $\mathbf{E} = \mathbf{e}$. From my perspective, allowing collections of variable values to be effects in this way does not purchase any additional generality; for I am inclined to say that $\mathbf{C} = \mathbf{c}$ caused $\mathbf{E} = \mathbf{e}$ iff $\mathbf{C} = \mathbf{c}$ caused $E_i = e_i$, for each $E_i \in \mathbf{E}$ and its corresponding value $e_i \in \mathbf{e}$.

Are joint causes causes simpliciter? Did Franny cause the proposal to pass? We could go either way. While the formalism will distinguish causes which are 1-tuples from causes which are n-tuples, for $n > 1$, we could decide to interpret this formalism by saying that if some n-tuple $\mathbf{C}$ caused E, then each $C \in \mathbf{C}$ counts as a cause of E in its own right. Or we could decide to say that each $C \in \mathbf{C}$ is merely *part* of a cause, and distinguish joint from individual causation. My own inclination is to say that neither Franny nor Sammy individually caused the proposal to pass, even though, together, they did; but if the reader balks at this, they should feel free to go the other way.

4. Preemptive Overdetermination

The neuron system shown in figure 1 provides a case of *Preemptive Overdetermination*. For another case with a similar causal structure, consider *Tax Cut*.[33]

> *Tax Cut*
> The proposal to lower corporate taxes requires one more vote to pass. Tammy's constituents will be angry if she votes in favor, but it is important to her campaign contributors that the proposal pass, so she is prepared to deal with her constituents' ire if her vote is needed. Fortunately for Tammy, Sammy votes 'yea', the proposal passes by a single vote, and Tammy is free to vote 'nay'.

The proposal's passing was overdetermined—the corporate donors bought more than enough influence. But the overdetermination is not symmetric. Though the causal process initiated with donations to Sammy runs to completion, the causal process initiated with donations to Tammy is preempted by Sammy's voting 'yea'. Tammy would have caused the proposal to pass, were it not for Sammy; but, as it happens, Tammy is merely a backup, would-be cause of the proposal's passing.

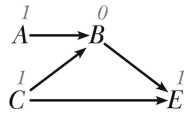
Cases like *Preemptive Overdetermination* serve as counterexamples to a simple counterfactual theory of causation which says that counterfactual dependence is necessary for causation. Consider the canonical model of the neuron system from figure 1, $\mathbf{M}_1$. In that model, it is not true that, had C not fired, E wouldn't have fired. For, had C not fired, B would have fired, and E would have fired all the same. (In the counterfactual model $\mathbf{M}_1[C \rightarrow 0]$ in which we intervene to set C's value to 0, E

33. When I say that *Tax Cut* has a similar causal structure, I assume that the corporate donations, the 'yea' votes, and the proposal's passing are all deviant.

takes on the value 1.) But C's firing caused E to fire. So counterfactual dependence is not necessary for causation.

Lewis (1973) dealt with cases like *Preemptive Overdetermination* by taking causation to be not counterfactual dependence but rather the ancestral, or the transitive closure, of counterfactual dependence. While E's firing doesn't counterfactually depend upon C's firing directly, it *does* counterfactually depend upon D's firing, and D's firing counterfactually depends upon C's firing. So Lewis says that C's firing caused E to fire. This Lewisian transitivity maneuver allows us to correctly say that, in the model $\mathbf{M}_1$, $C = 1$ caused $E = 1$. Unfortunately, if we straightforwardly import the Lewisian maneuver into the framework of causal models, the resulting account will be model-variant. For suppose we remove the variable D from $\mathbf{M}_1$, in the manner described in section 2.2. We will get the model $\mathbf{M}_1^{-D}$, in which there is no variable intermediate between C and E.

$$E := B \vee C$$
$$B := A \wedge \neg C$$



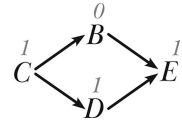
Even though, given the causal model $\mathbf{M}_1$, a Lewisian theory will say that $C = 1$ caused $E = 1$, given the model $\mathbf{M}_1^{-D}$, it will say that $C = 1$ didn't cause $E = 1$. So the theory will be model-variant.

The treatment of *Preemptive Overdetermination* favored by almost every author in the causal modeling literature appeals to either A or B.[34] Though $E = 1$ does not counterfactually depend upon $C = 1$ in the model $\mathbf{M}_1$, it *does* counterfactually depend upon $C = 1$ in the *counterfactual* model where we've intervened to fix B's value to 0: $\mathbf{M}_1[B \rightarrow 0] \vDash C = 0 \,\square\!\!\rightarrow\, E = 0$. Likewise, $E = 1$ counterfactually depends upon $C = 1$ in the counterfactual model $\mathbf{M}_1[A \rightarrow 0]$. And according to these authors, counterfactual dependence in counterfactual models like these is sufficient to show that $C = 1$ caused $E = 1$. No solution which appeals to the variables A or B in this way will be model-invariant. For note that the exogenous variable A is inessential in $\mathbf{M}_1$. So, by **Exogenous**

34. See, in particular, Halpern and Pearl 2001, 2005; Hitchcock 2001; Woodward 2003; Halpern 2008, 2016; Weslake, forthcoming. See Yablo 2002, 2004 for similar ideas. Andreas and Günther (2018) have a different treatment of *Preemptive Overdetermination* which also appeals to the variable B. (Beckers and Vennekens [2017, 2018] have a *radically* different treatment of *Preemptive Overdetermination*—according to them, preemptive overdeterminers are not causes.)

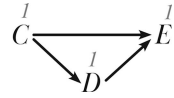
Removal, we may pluck it out, and we will be left with a model, $\mathbf{M}_1^{-A}$, in which the endogenous variable B is (now) inessential.

$$E := B \vee D$$
$$D := C$$
$$B := \neg C$$



Since B is inessential, **Endogenous Removal** tells us that we may pluck it out. Doing so leaves us with a model, $\mathbf{M}_1^{-A,-B}$, in which neither A nor B appears.

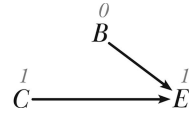
$$E := \neg C \vee D$$
$$D := C$$



So, if we want our theory of causation to be model-invariant, then we will want a treatment of *Preemptive Overdetermination* which does not require the variables A or B.

Return to the causal model $\mathbf{M}_1^{-D}$. For a moment, ignore the structural equation for B, focus just on E's structural equation, and treat this isolated structural equation as if it were a causal model unto itself—what we can call the *local* model at E.

$$E := B \vee C$$



Notice that, in the local model at E, there will be counterfactual dependence between $E = 1$ and $C = 1$. Since this is so, I'll say that $E = 1$ *locally* counterfactually depends upon $C = 1$.

In general, given a causal model $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, with $E \in \mathbf{V}$, let's define the *local model at E*, which we can write '$\mathbf{M}(E)$', to be the model in which (a) the exogenous variables are just the parents of E, $\mathbf{PA}(E)$, in the original model $\mathbf{M}$; (b) these exogenous variables are assigned whatever values they take on in $\mathbf{M}$; (c) the sole endogenous variable is E; (d) the sole structural equation is E's structural equation from $\mathbf{M}$; and (e) the information about the deviancy of E and $\mathbf{PA}(E)$'s values is the same as in $\mathbf{M}$. Then, we may say that, in the model $\mathbf{M}$, $E = e$, rather than e^*, *locally* counterfactually depends upon $C = c$, rather than c^*, iff, in the local

model at E, $E = e$, rather than e^*, counterfactually depends upon $C = c$, rather than c^*:

$$\mathbf{M}(E) \vDash C = c^* \,\square\!\!\rightarrow E = e^*$$

In contrast, if there is counterfactual dependence in the model $\mathbf{M}$,

$$\mathbf{M} \vDash C = c^* \,\square\!\!\rightarrow E = e^*$$

Then I will say that $E = e$, rather than e^*, *globally* counterfactually depends upon $C = c$, rather than c^*, in the model $\mathbf{M}$.[35] (If C is a causal parent of E and there is only one path leading from C to E, then there won't be any difference between local and global dependence—in those cases, I will allow myself to say simply: '$E = e$, rather than e^*, *depends* upon $C = c$, rather than c^*'.)

To properly classify $C = 1$ as a cause of $E = 1$ in $\mathbf{M}_1^{-D}$, I will suggest that we focus on *local*, as opposed to *global*, counterfactual dependence. Turning our attention to local dependence may help with $\mathbf{M}_1^{-D}$, but it will not, on its own, help us to say that C's firing caused E to fire in the canonical model $\mathbf{M}_1$. For in this model, E's firing does not locally depend upon C's firing (the variable for C is not even included in the local model $\mathbf{M}_1(E)$). I believe that we should handle this case roughly as Lewis (1973) did: by focusing not on local dependence but rather on something like the *transitive closure* of local dependence. However, there are a number of counterexamples to the thesis that a chain of dependence is sufficient for causation. Let's turn to those counterexamples now.

5. Causal Networks

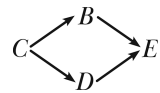
Suppose you've traced out a sequence of states or events, where each state or event in the sequence depends upon its predecessor. When can you go on to conclude that the state or event at the start is a cause of the one at the end? Lewis gave the answer 'always'. This answer allowed him to deal with cases like *Preemptive Overdetermination*, but it came at a cost. Chris smokes, contracts cancer, undergoes chemo, and survives. The survival depends upon the chemo; the chemo depends upon the cancer; and the cancer depends upon the smoking. Lewis concludes that smoking caused Chris to survive. This is difficult to swallow, no matter how it's seasoned.

35. Of course, in order for these dependence claims to be true, it must also be that $\mathbf{M} \vDash C = c \wedge E = e$. Throughout, I am using 'c' and 'e' for the actual values of C and E. I will say more about the contrastive 'rather than' clauses in section 5.1 below.

The answer to give is 'sometimes, but not always', and the difficulty lies in working out just when.

In this section, I will try to lay down conditions specifying when a directed path running from C to E, $\mathcal{P}: C \rightarrow D_1 \rightarrow D_2 \rightarrow \ldots \rightarrow D_N \rightarrow E$, is what I will call a *causal path*. Actually, I will try to do something slightly more general. In section 3, I explained that I will provide a theory of causation which allows *tuples* of variable values to be causes. But there won't be a *single* directed path from a tuple of variables **C** to an effect variable E. So I will begin by generalizing the notion of a directed path—I'll call the generalization a *network*—and then I'll try to lay down conditions specifying when a *network* from **C** to E is what I will call a *causal network*. My theory will say that causal networks are necessary for causation: if **C**'s values are to be a cause of E's, then there must be a causal network leading from **C** to E.

First, let me explain what I mean by *network*. We may think of a directed path, $\mathcal{P}$, from C to E, as a collection of directed edges generated by the following procedure: begin with C, and select exactly one of its causal children, D, to be its $\mathcal{P}$-child. Then, include the directed edge between C and D, $C \rightarrow D$, in $\mathcal{P}$. Next, select exactly one of D's causal children, and proceed in this manner until you reach E. Now, we can define a *network*, $\mathcal{N}$, from the tuple of variables **C** to E, as a collection of directed edges generated by the following procedure: begin with each variable $C \in \mathbf{C}$, and select some of its causal children, $D_1, D_2, \ldots, D_N$ (you needn't choose just one), to be its $\mathcal{N}$-children.[36] Next, for each of the D_i, select some of their causal children to be their $\mathcal{N}$-children, and proceed in this manner until E is the only variable in $\mathcal{N}$ without an $\mathcal{N}$-child. That is, a *network* from **C** to E is just a union of directed paths from some $C \in \mathbf{C}$ to E—where, for each $C \in \mathbf{C}$, there is some directed path leading from C to E included in the union. For instance, in $\mathbf{M}_6$, $A \rightarrow E \leftarrow C$ is a network from (A, C) to E. And, in $\mathbf{M}_1$,



36. Terminology: if there is a directed edge $C \rightarrow D$ in a network $\mathcal{N}$, then I say that D is one of C's $\mathcal{N}$-children, and that C is one of D's $\mathcal{N}$-parents. Note that being one of D's $\mathcal{N}$-parents is not the same as being a parent of D lying in the network $\mathcal{N}$. Consider the network $\mathcal{N}: C \rightarrow B \rightarrow E$ in the model $\mathbf{M}_1^{-D}$. C is a parent of E lying in $\mathcal{N}$, but C is not one of E's $\mathcal{N}$-parents.

is a network from C to E (remember, I don't distinguish between the variable C and the 1-tuple (C)). Note that every directed path is a network, though not every network is a directed path.

To reiterate: in this section, I will be trying to lay down conditions specifying when a *network* is causal. And according to the theory I'll present in section 6, causal networks are necessary for causation. In order for **C**'s values to be a cause of E's value, there must be a causal network leading from **C** to E. In these terms, a Lewisian view says that a network $\mathcal{N}$ is causal whenever the value of each variable in $\mathcal{N}$ depends upon the values of its $\mathcal{N}$-parents. I believe that we should impose additional constraints on a network being causal. I'll introduce these constraints by surveying some representative counterexamples to this Lewisian view.

5.1. Causal Networks and Contrasts

One class of counterexamples to the Lewisian view is well illustrated by the neuron system illustrated in figure 7 (cf. Paul and Hall 2013: figure 17; Lewis 1986: 210).

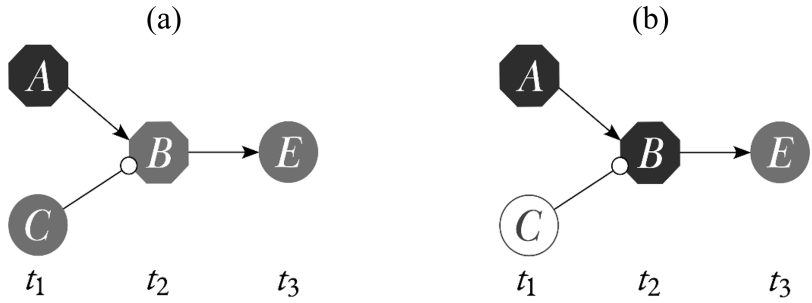


Figure 7.

In this neuron system, the octagonal neurons A and B are special. They can either fire *weakly* (indicated with light gray coloring) or *strongly* (indicated with dark gray). The connection between C and B is a special kind of inhibitory connection—if the neuron at its base fires, then this will diminish the strength with which the neuron at its head would otherwise have fired. So, for example, if A fires strongly and C doesn't fire, as in figure 7b, then B will fire strongly. But if A fires strongly and C fires, as in figure 7a, then B will only fire weakly. Neuron E is a regular neuron, so if B fires, whether weakly or strongly, E will fire. In figure 7a, E's firing (rather than not) depends upon B's firing weakly (rather than not firing). And

B's firing weakly (rather than strongly) depends upon *C*'s firing (rather than not). But *C*'s firing did not cause *E* to fire. So this neuron system provides a counterexample to the Lewisian view that causation is the transitive closure of dependence.

For another case with a similar structure: A dog bites Michael's right hand. With his right hand on the mend, Michael uses his left hand to hail a taxi. The taxi's stopping depends upon Michael's hailing the taxi with his left hand (rather than not hailing the taxi), and Michael's hailing the taxi with his left hand (rather than his right) depends upon the dog bite. But the dog bite did not cause the taxi to stop.[37]

I follow Cei Maslen (2004) and Jonathan Schaffer (2005) in thinking that cases like these illustrate the importance of paying attention to *contrasts* in chains of dependence.[38] There is a difference between saying that (a) $E = e$, rather than e^*, depends upon $C = c$, rather than c^*, and saying that (b) $E = e$, rather than e^{**}, depends upon $C = c$, rather than c^*, or that (c) $E = e$, rather than e^*, depends upon $C = c$, rather than c^{**}. The first claim, (a), is made true by a counterfactual $C = c^* \Box\!\!\rightarrow E = e^*$; the second, (b), is made true by a counterfactual $C = c^* \Box\!\!\rightarrow E = e^{**}$; and the third, (c), is made true by a counterfactual $C = c^{**} \Box\!\!\rightarrow E = e^*$. The lesson of figure 7 is this: in order for a network to be causal, it is not enough that the value of each variable in the network depend upon the value of its parents in the network. The relevant contrasts also have to 'match up'.

As a preliminary account, then, we have:

> Causal Network (preliminary)
> A network, $\mathcal{N}$, from **C** to *E*, is a *causal network* only if there is an assignment of contrasts to the variables in $\mathcal{N}$ such that:
> (a) *E*'s contrast is distinct from its value;
> (b) for each $D \notin \mathbf{C}$ in the network, *D*'s value, rather than its contrast, locally depends upon *D*'s $\mathcal{N}$-parents' values, rather than their contrasts.

And our preliminary theory is that $\mathbf{C} = \mathbf{c}$ caused $E = e$ only if there is a causal network leading from **C** to *E*. Note that there is no *one* contrast we could assign *B* in figure 7a such that *E*'s firing, rather than not, depends

37. See McDermott 1995, as well as the counterexamples to transitivity discussed in Paul 2004.

38. For more on contrasts in causal claims, see Hitchcock 1996a, 1996b; Schaffer 2012a.

upon B's firing weakly, rather than that contrast; and such that B's firing weakly, rather than that contrast, depends upon C's firing, rather than not. So $C \rightarrow B \rightarrow E$ is not a causal network, and our preliminary theory tells us that C's firing was not a cause of E's firing.

Note that, because we require the contrasts to 'match up', once we have chosen contrasts for the variables in $\mathbf{C}$, the choice of every other contrast is out of our hands. Pick any $D \notin \mathbf{C}$ in the network $\mathcal{N}$, let $\mathbf{P}$ be its $\mathcal{N}$-parents, and let $\mathbf{p}^*$ be their contrasts. Then, clause (b) tells us that D's contrast must be the value d^* such that $\mathbf{P} = \mathbf{p}^* \boxdot\!\!\rightarrow D = d^*$ is true in the local model at D. There will only be one such d^*, so we have no choice about which contrast to assign to D. (D was arbitrary, save our assumption that $D \notin \mathbf{C}$, so the same goes for every variable in the network, except for those in $\mathbf{C}$.)

Paying attention to contrasts has other benefits as well. For instance, it allows us to handle cases of *trumping preemption* (see Schaffer 2004). Suppose that the troops always follow the orders of the highest ranked officer. The Major and the Sergeant both order the troops to advance, and they advance. Since the Major outranks the Sergeant, it is natural to want to say that it was the Major, and not the Sergeant, who caused the troops to advance. Use a variable, M, to represent the Major's orders. Let M take on the value 2 if the Major orders to advance, 1 if he orders to stay put, and 0 if he gives no order at all. Similarly, use the variable S for the Sergeant's orders. S is 2 if the Sergeant orders to advance, 1 if he orders to stay put, and 0 if he gives no orders at all. And, finally, use a variable, A, for whether the troops advance. $A = 2$ if they advance, and $A = 1$ if they do not. I'll assume that the structural equation $A := \phi_A(M, S)$ is correct, where

$$\phi_A(M, S) = \begin{cases} M, \text{ if } M \neq 0 \\ S, \text{ if } M = 0 \text{ and } S \neq 0 \\ 1, \text{ if } M = 0 \text{ and } S = 0 \end{cases}$$

That is: the soldiers will do whatever the Major orders, so long as the Major gives an order. If he does not, then they will follow the orders of the Sergeant. If neither the Major nor the Sergeant give orders, then they will not advance. In this model, notice that, even though the soldiers' advance doesn't depend upon the Major's giving the order to advance, rather than giving no orders at all ($M = 0 \boxdot\!\!\rightarrow A = 2$), it *does* depend upon the Major's giving the order to advance, rather than giving the order to stay put ($M = 1 \boxdot\!\!\rightarrow A \neq 2$). So $M \rightarrow A$ will be a causal network.

Since the soldiers' advance does not depend upon the Sergeant's orders, no matter which contrast we choose, $S \rightarrow A$ will not be a causal network, and the Sergeant's orders will not count as a cause of the soldiers' advance.[39]

5.2. Causal Networks, Defaults, and Deviancy

Schaffer (2005) holds that this kind of contrastivism allows us to handle all counterexamples to the Lewisian view, but in the present context, this would be an overreach.[40] Consider again the neuron system of *Preemptive Overdetermination* from figure 1, but suppose that C doesn't fire, as in figure 8. In this neuron system, E's firing depends upon B's firing (rather than not). And B's firing (rather than not) depends upon C's dormancy. So we have a chain of dependence with matching contrasts leading from C to E, but C's value did not cause E's.[41]

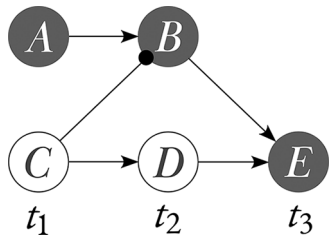


Figure 8.

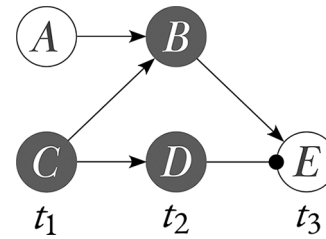


Figure 2.

Or consider again the neuron system from figure 2 (reproduced here). There, E's remaining dormant depends upon D's firing (rather than not), and D's firing (rather than not) depends upon C's firing. So again we have a chain of dependence with matching contrasts leading from C to E, but C's value did not cause E's.

As we've already seen (in section 1.1), were it not for the information about which variable values are default, inertial states and which are deviant noninertial events, we could model the neuron system in

39. Cf. the treatments of *trumping preemption* in Lewis 2004; Halpern and Hitchcock 2010; Hitchcock 2011.

40. Schaffer is working in a different theoretical framework, and it affords him a response to the kinds of counterexamples raised below (see Schaffer 2005: 342).

41. Carolina Sartorio's *Causes as Difference Makers* principle (2005, 2016) entails that C's failure to fire cannot cause E to fire, so long as C's firing would have caused E to fire.

figure 2 with a model isomorphic to the canonical model of *Preemptive Overdetermination* from figure 1. So we should expect an explanation of why $C = 1$ didn't cause $E = 0$ to make use of this additional information. Note also that **Exogenous Removal** and **Endogenous Removal** allow us to remove every variable other than C and E from $\mathbf{M}_2$. A is inessential, so **Exogenous Removal** tells us that the model $\mathbf{M}_2^{-A}$ is correct. In the model $\mathbf{M}_2^{-A}$, B is inessential, so **Endogenous Removal** tells us that the model $\mathbf{M}_2^{-A,-B}$ is correct. And similarly, in the model $\mathbf{M}_2^{-A}$, D is inessential, so **Endogenous Removal** tells us that the model $\mathbf{M}_2^{-A,-D}$ is correct. If we want our theory of causation to be model-invariant, then it had better tell us that $C = 1$ didn't cause $E = 0$ in each of these models. So we have good reason to think that the verdicts of our theory should not depend upon the default information of any variables other than C and E themselves.[42]

In both figure 2 and figure 8, it is noteworthy that either C or E takes on a value representing a default, normal, or inertial state. Whereas, in figure 1, both C and E take on values representing deviant, abnormal, noninertial events. It is also noteworthy that, in both $\mathbf{M}_2$ and $\mathbf{M}_8$, there are multiple directed paths from C to E. I will suggest that these are the reasons why C does not cause E in either of these neuron systems.

Suppose that we are given a network, $\mathcal{N}$, from $\mathbf{C}$ to E, and in this network are two variables, D and R. If there is a directed path from D to R, $\mathcal{O} : D \rightarrow O_1 \rightarrow O_2 \rightarrow \ldots \rightarrow O_N \rightarrow R$, where none of the directed edges in $\mathcal{O}$ are included in $\mathcal{N}$, then I'll say that D is a *departure* variable, and that R is one of its *return* variables (relative to the network $\mathcal{N}$). For instance, in the model $\mathbf{M}_8$, relative to the network $C \rightarrow B \rightarrow E$, C is a departure variable and E is its return. And, in the model $\mathbf{M}_2$, relative to the network $C \rightarrow D \rightarrow E$, C is a departure variable with return E. In contrast, relative to the network $A \rightarrow B \rightarrow E$ in $\mathbf{M}_2$, E is *not* a return variable—and, relative to the network $C \rightarrow B \rightarrow E \leftarrow D \leftarrow C$, C is not a departure variable.

Take some network, $\mathcal{N}$, with a departure variable D, and one of its returns, R. D potentially affects R both via $\mathcal{N}$ and via some other path or paths external to $\mathcal{N}$. It could be that what D gives R through $\mathcal{N}$, it takes away along some other path or paths. If D gives a deviant value to R through $\mathcal{N}$—that is, if both D and R have deviant values and more

42. Every variable in the model besides C and E may be removed; but we may not remove every variable besides C and E. For D is not inessential in $\mathbf{M}_2^{-A,-B}$, and B is not inessential in $\mathbf{M}_2^{-A,-D}$. So for all we've said, it could be that what $\geq$ tells us about D should be relevant to the theory's verdicts in $\mathbf{M}_2^{-A,-B}$, while what $\geq$ tells us about B should be relevant to the theory's verdicts in $\mathbf{M}_2^{-A,-D}$.

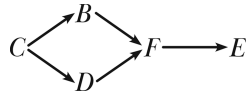
default contrasts—then this will make no difference with respect to whether $\mathcal{N}$ is a causal network. (Thus, in figure 1, $C \rightarrow D \rightarrow E$ is causal.) But if D does not give a deviant value to R through $\mathcal{N}$, then $\mathcal{N}$ is not a causal network. (Thus, in figure 2, $C \rightarrow D \rightarrow E$ is not causal.)

Let us add this to our account: a network is causal only if every departure and return variable in the network takes on a value which is more deviant than its contrast.[43]

> CAUSAL NETWORK
> A network, $\mathcal{N}$, from **C** to E, is a *causal network* iff there is an assignment of contrasts to the variables in $\mathcal{N}$ such that:
> (a) E's contrast is distinct from its value;
> (b) for each $D \notin \mathbf{C}$ in the network, D's value, rather than its contrast, locally depends upon D's $\mathcal{N}$-parents' values, rather than their contrasts; [44] and
> (c) every departure and return variable in $\mathcal{N}$ has a value which is more deviant than its contrast.

This completes my account of when a network is causal.

Note that, while CAUSAL NETWORK requires E's contrast to be distinct from its value, it does not require that the *other* variables in the network have contrasts which are distinct from their values.[45] For instance, consider the neuron system in figure 9. This is a case of *double prevention*. F is a potential preventer of E's firing; and C's firing prevented F from preventing E. In the canonical model $\mathbf{M}_9$,



is a causal network from C to E. For we may assign C, B, D, and E the contrast value 0 (note that B's contrast is the same as its value) and F the contrast value 1. Then, $E = 1$, rather than 0, locally depends upon

43. Couldn't a departure variable, D, have a value no more deviant than its contrast, and yet still not take away along other paths what it gives to its return variable, R, through $\mathcal{N}$? Yes, but in that case, the additional paths from D to R may simply be incorporated into the network $\mathcal{N}$, and the resulting network will be causal. (See the discussion of figure 9 below.)

44. Recall: there is a difference between a variable's $\mathcal{N}$-parents and its causal parents lying in $\mathcal{N}$. See note 36.

45. If d^* is D's actual value, it is odd to call d^* a *contrast* value, but I'll stick to this terminology nonetheless.

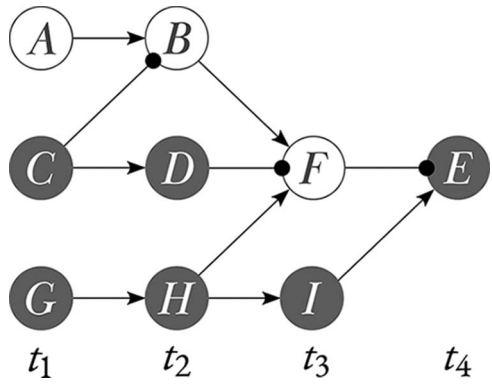


Figure 9.

$F = 0$, rather than 1. $F = 0$, rather than 1, locally depends upon $(B, D) = (0, 1)$, rather than $(0, 0)$. $D = 1$, rather than 0, locally depends upon $C = 1$, rather than 0. And $B = 0$, rather than 0, locally depends upon $C = 1$, rather than 0. (For, in the local model at B, $\mathbf{M}_9(B)$, the counterfactual $C = 0 \mathrel{\square\!\!\rightarrow} B = 0$ is true.) It can seem that the variable B is an idle wheel in this network, but it is important that it be included. For, relative to the network $C \rightarrow D \rightarrow F \rightarrow E$, F is a return variable with a default value and a deviant contrast. So the network $C \rightarrow D \rightarrow F \rightarrow E$ is not causal. However, relative to the network which includes B, F is not a return variable, and need not have a deviant value, nor a default contrast.

Note that E's firing globally counterfactually depends upon C's firing. If we think that global counterfactual dependence between events like these suffices for causation, and we wish to understand causation in terms of causal networks, then it is for the good that we count as causal the network which includes B. In fact, global counterfactual dependence suffices for the existence of a causal network, not just for the model $\mathbf{M}_9$, but *in general*. That is: in any causal model $\mathbf{M}$, if there is some assignment $\mathbf{c}^*$ to the variables in $\mathbf{C}$ such that the global counterfactual $\mathbf{C} = \mathbf{c}^* \mathrel{\square\!\!\rightarrow} E \neq e$ is true, then there will be a causal network leading from some subtuple of $\mathbf{C}$ to E in $\mathbf{M}$. (See proposition A.1 in the appendix for a proof.)

So defined, causal networks are model-invariant. Suppose we have a causal model $\mathbf{M}$, with an inessential exogenous variable $U \notin \mathbf{C}$. Then, there will be a causal network from $\mathbf{C}$ to E in $\mathbf{M}$ iff there is a causal network from $\mathbf{C}$ to E in $\mathbf{M}^{-U}$. Similarly, if we have a causal model $\mathbf{M}$, with an inessential endogenous variable $V \notin \mathbf{C} \cup (E)$, then there will be a causal network from $\mathbf{C}$ to E in $\mathbf{M}$ iff there is a causal network from $\mathbf{C}$ to E in $\mathbf{M}^{-V}$. (See the proof of proposition A.2 in the appendix.)

If we suppose that survival is an inertial state—the state in which people normally remain unless they are acted upon from without—then this proposal explains why the boulder's becoming dislodged does not cause Matthew to survive (in *Boulder*, from section 1.1), even though his survival depends upon his jumping out of the way (rather than staying put), and his jumping out of the way (rather than his staying put) depends upon the boulder's getting dislodged. So too does it explain why Chris's smoking does not cause him to survive, even though his survival depends upon the chemotherapy, and the chemotherapy depends upon the smoking. Both cases have a causal structure similar to *Short Circuit*: a threat to survival is created along one path, and simultaneously neutralized along another. If survival is an inertial state, then neither path will be causal. (Nor will the network which consists of *both* paths be causal—for, while the survival depends upon the neutralization of the threat, it does not depend upon the threat and the neutralization both. If Chris had neither cancer nor chemo, he would still have survived; and, had the boulder not fallen and Matthew not jumped, he would still have survived.)

6. Causation and the Transmission of Deviancy

Causal networks are the model-invariant heart of my theory of causation. On my view, in order for **C** to cause *E*, there must be a causal network leading from **C** to *E*. In section 6.1, I'll say a bit to motivate thinking of a causal network as a process which transmits deviant, abnormal, or noninertial behavior. In section 6.2, I'll present my theory of causation, according to which (roughly) **C** is a cause of *E* iff there is a causal network leading from **C** to *E*, **C** has deviancy to give, and *E* receives that deviancy via the causal network. I'll go on to apply this theory to cases from Sarah McGrath (2005) and Ned Hall (2004).

6.1. Productive Networks

The distinction between the values of variables which represent default, normal, inertial states and those which represent deviant, abnormal, noninertial events enters into my theory of causation at least in clause (c) of CAUSAL NETWORK. It is natural to wonder what this distinction is doing in a theory of causation. I take the argument presented in section 1.1 to demonstrate that this distinction or something like it *must* be included in any adequate theory. But, even once this is appreciated, it is natural to wonder: *why* should this distinction play any role in our causal thought and

talk? In this subsection, I want to gesture at an answer to this question. Roughly, I will suggest that a cause is something which transmits abnormal, deviant, or noninertial behavior to its effect.

If causation is to be understood in terms of the transmission of deviancy, then what is it for this deviancy to be *transmitted*? One possible answer is that deviancy is transmitted iff there is an uninterrupted process leading from cause to effect, each stage of which receives its deviancy from the preceding stage. Let's try to make this idea a bit more precise. Contrast a *causal* network, as defined in section 5, with a *productive* network, as defined below. (The only difference is in clause (c).)

> PRODUCTIVE NETWORK
> A network, $\mathcal{N}$, from $\mathbf{C}$ to E, is a *productive network* iff there is an assignment of contrasts to the variables in $\mathcal{N}$ such that:
> (a) E's contrast is distinct from its value;[46]
> (b) for each $D \notin \mathbf{C}$ in the network, D's value, rather than its contrast, locally depends upon D's $\mathcal{N}$-parents' values, rather than their contrasts; and
> (c) every variable in $\mathcal{N}$ has a value which is more deviant than its contrast.

Note that any productive network will automatically count as a *causal* network. But not all causal networks are productive networks. Being linked by a productive network is sufficient, but not necessary, for being linked by a causal network.

A productive network is so called because it provides a natural characterization of the notion of a productive causal process in terms of causal models.[47] So understood, a productive causal process is an uninterrupted process by which deviant values are transmitted. And what it is for this deviancy to be *transmitted* is for the deviancy of each stage to locally depend upon the deviancy of its immediate predecessors.

Notice that there is a productive network leading from C to E in the canonical model of *Preemptive Overdetermination* in figure 1. So too is there a productive network leading from A to E in the canonical models of figures 4, 7, and 8—and from G to E in figure 9. In general, it seems

46. Condition (a) is redundant in the presence of condition (c), but I include it to emphasize that PRODUCTIVE NETWORK is a strengthening of CAUSAL NETWORK.

47. The notion which PRODUCTIVE NETWORK characterizes is not the notion of a causal process provided by authors like David Fair (1979), Wesley Salmon (1984, 1994), and Phil Dowe (2000)—those notions are characterized in terms of physics, not causal models—but there are some similarities. Cf. also Hall's (2004) characterization of causal production.

that, if there is a productive network from C to E in the canonical model, the judgment that C caused E is intuitive and uncontroversial. There is little debate about whether C's firing caused E to fire in figure 1, or whether A's firing caused E to fire in figures 4, 7, and 8. In contrast, in the canonical model of the case of *Double Prevention* shown in figure 10, there is a *causal*, but not a *productive*, network leading from C's firing to E's firing. In $\mathbf{M}_{10}$, $C \rightarrow D \rightarrow E$ is a causal network. However, $C \rightarrow D \rightarrow E$ is not a productive network, since the intermediate variable D takes on a default value. People's causal judgments about figure 10 tend to be less uniform. More generally, it seems that, when variables are connected by causal, but not productive, networks, some (but by no means all) are more hesitant to attribute causation.

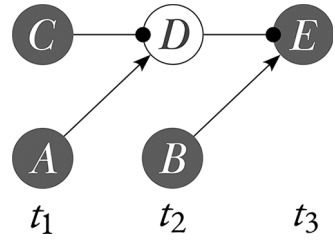
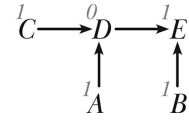


Figure 10. *Double Prevention.*

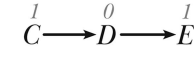
Unlike causal networks, *productive* networks are model-variant. Take the canonical model $\mathbf{M}_{10}$.

$$E := B \wedge \neg D$$
$$D := A \wedge \neg C$$



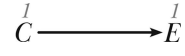
In this model, the exogenous variables A and B are both inessential. So **Exogenous Removal** tells us that we may remove them both, leaving behind the model $\mathbf{M}_{10}^{-A,-B}$.

$$E := \neg D$$
$$D := \neg C$$



In this model, the exogenous variable D is inessential, so **Endogenous Removal** tells us that we may remove it, leaving behind the model $\mathbf{M}_{10}^{-A,-B,-D}$.

$$E := C$$



And in *this* model, there is a productive network leading from C to E.

So: if we were to understand the transmission of deviancy as PRODUCTIVE NETWORK does—each variable intermediate between $\mathbf{C}$ and E receives a deviant value from its parents in the network—then whether deviancy is transmitted from $\mathbf{C}$ to E will vary from model to model.[48] CAUSAL NETWORK is a model-invariant weakening of PRODUCTIVE NETWORK. It suggests a different way of understanding the transmission of deviancy. Suppose that $E = e$, rather than e^*, globally counterfactually depends upon $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, and suppose that $\mathbf{C}$ and E both represent deviant, noninertial events, while both $\mathbf{c}^*$ and e^* represent more default, inertial states. In that case, let us say that $\mathbf{C}$ has transmitted deviancy to E—we won't concern ourselves with whether, for instance, this transmission was accomplished with double prevention or not. Because global counterfactual dependence suffices for the existence of a causal network, if E's deviancy counterfactually depends upon $\mathbf{C}$'s, then there will be a causal network leading from (some subtuple of) $\mathbf{C}$ to E. Moreover, if E globally depends on $\mathbf{C}$, there will be a causal network leading from (some subtuple of) $\mathbf{C}$ to E *without* any departure or return variables—call this a 'closed causal network'.[49] So another, equivalent, way of understanding the claim that counterfactual dependence between deviant, noninertial events suffices for the transmission of deviancy is this: deviancy may be transmitted through a closed causal network.

In the case of *Preemptive Overdetermination* from figure 1, E's deviancy does not globally depend upon C's. This is because C affects E along two separate paths. Along one path, C deprives E of deviancy; along the other, it provides deviancy. In cases like these, too, let us say that deviancy has been transmitted from cause to effect. More generally, if there are departure and return variables in a network, D and R, then it may be that what D transmits to R via the network, it takes away along some other path or paths. If D transmits *deviancy* to R through the network (if both D and R take on deviant, rather than more default, values), then this won't matter. We should still say that $\mathbf{C}$ has transmitted deviancy

48. Schaffer (2000, 2012b) argues that, in many paradigm instances of productive causal processes—pulling the trigger, thereby shooting the gun, thereby killing the target—we may interpolate variables between cause and effect so as to reveal a case of double prevention.

49. See the proof of proposition A.1 in the appendix to understand why, if $E = e$ counterfactually depends upon $\mathbf{C} = c$, there will be a closed causal network leading from (some subtuple of) $\mathbf{C}$ to E.

to *E*. That is: in general, we should allow deviancy to be transmitted through *any* causal network, and not just closed causal networks.

6.2. Productive Causation

CAUSAL NETWORK does not say anything about **C** and *E* having deviant values or (more) default contrasts. So if we wish to think of causation in terms of the transmission of deviancy in the way that I have been suggesting, then we should impose this additional requirement. Doing so yields the following relation, which I will call *productive causation*:

> PRODUCTIVE CAUSATION
> Given a causal model **M** containing the variables in **C** and *E*, $\mathbf{C} = \mathbf{c}$ is a *productive cause* of $E = e$ in **M** iff, in **M**, there is a minimal causal network leading from **C** to *E* which assigns contrasts to **C** and *E* which are more default than their values.

That is: $\mathbf{C} = \mathbf{c}$ is a productive cause of $E = e$ iff there's a minimal causal network leading from **C** to *E* and, *additionally*, **C** and *E*, like any departure and return variables in the network, have values which are more deviant than their contrasts. (I'll explain what I mean by 'minimal' below.)

If causation just is productive causation, this would explain some otherwise puzzling features of our causal thought and talk. To borrow an example from McGrath (2005): Alice's neighbor Bob promises Alice that he will water her plant while she is away on vacation. He doesn't, and Alice's plant dies. Many judge that Bob's failure to water the plant caused it to die. Only philosophers in the grip of theory judge that Alice's other neighbor, Carlos, caused the plant to die—though the plant's death counterfactually depends upon Carlos's failure to water it every bit as much as it depends upon Bob's failure to water it.[50] If we suppose that death and promise breaking are both deviant events, and that survival and promise-keeping are (more) default, then Bob's failure to water the plant is a productive cause of its death. And if we suppose that Carlos's failure to water the plant is a default state, then Carlos's failure to water is not a productive cause of its death.

If causation is productive causation, this allows us to explain why *switches* are not causes (see Hall 2004 and Sartorio 2005). For, while switches affect the *route* by which deviancy is transmitted to an effect, they do not themselves transmit deviancy to the effect.

50. See also the pen case in Hitchcock and Knobe 2009.

For a concrete case of a switch, consider the neuron system shown in figure 11a. There, the neuron *S* is a *switch*, which can either be set *left* (when the variable *S* is even, as in figures 11b and 11d) or *right* (when the variable *S* is odd, as in figures 11a and 11c). *D* determines whether the switch is set left or right. If *D* fires, then *S* will be set right; whereas, if *D* does not fire, then *S* will be set left. *D* does not determine whether *S* fires or not. *M* does that. If *M* fires, then *S* will fire; if *M* does not fire, then *S* will not fire. If *S* fires while left, then *L* will fire. If *S* fires while right, then *R* will fire. And, finally, *E* will fire iff either *L* or *R* does.

For a case with a similar structure, consider:

> *Doorbells*
> There are two doorbells—one on the left, and one on the right. The signal from the button outside passes through a switch, which can have one of two settings: left or right. If the switch is set to the left and the button is pressed, the signal will pass to the left, and the left bell will ring. If the switch is set to the right and the button is pressed, the signal will pass to

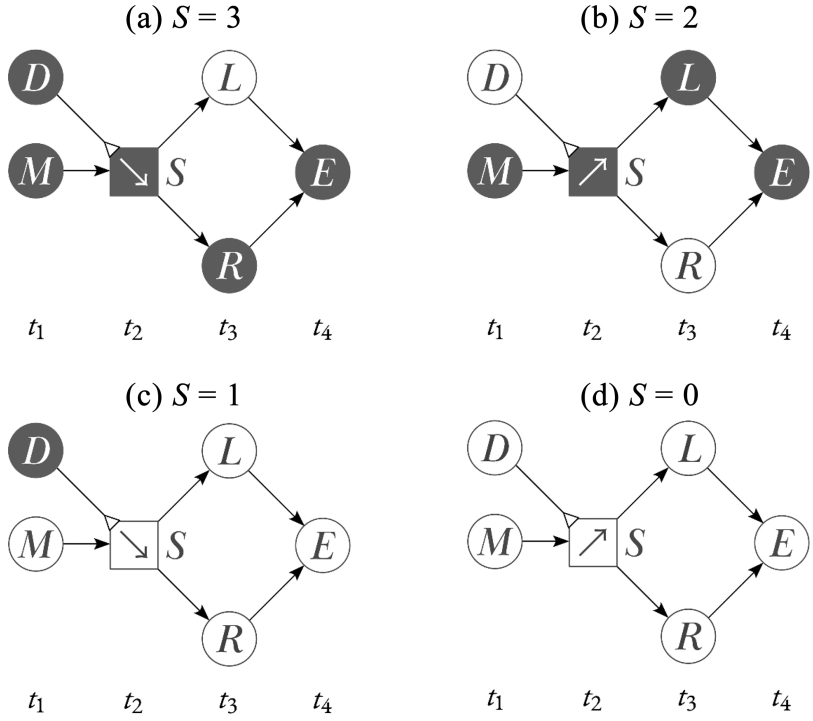


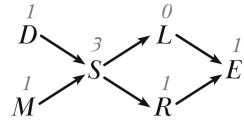
Figure 11. *Switch.* The neuron *S* can either be set to the left or to the right. If *D* fires, then it will be set to the right; if *D* doesn't fire, then it will be set to the left. *S* will fire iff *M* fires.

> the right, and the right bell will ring. If either bell rings, Einstein will bark. Before leaving that morning, Doc flipped the switch to the right. When Marty arrives, he presses the button, the right bell rings, and Einstein barks.

In *Doorbells*, when Marty presses the button, Einstein will bark—no matter whether the switch is set to the left or the right. Doc's flipping the switch to the right was (along with Marty's pressing the button) a cause of the right bell's ringing, but it was not a cause of Einstein's barking.[51] In contrast, Marty's pressing the button *was* a cause of Einstein's barking. Likewise, in figure 11a, while *D*'s firing was a cause of *R*'s firing, it was not a cause of *E*'s firing. In contrast, *M*'s firing *was* a cause of *E*'s firing.

I'll assume that both of these systems can be modeled with the following system of structural equations.[52]

$$E := L \vee R$$
$$L := S = 2$$
$$R := S = 3$$
$$S := 2M + D$$



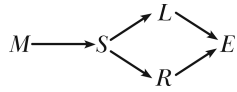
I will also assume that $S = 2$ is no more deviant or abnormal than $S = 3$—being set to the left is no less normal than being set to the right. With this assumption, we can show that, while there is a causal network from *M* to *E*, there is no causal network from *D* to *E*.

First, let's assume that $S = 3$ is more deviant than $S = 1$ and that $E = 1$ is more deviant than $E = 0$—in the case of *Switch*, firing is more deviant than remaining dormant, or, in the case of *Doorbells*, directing a signal right is more deviant than not directing any signal, and barking is more deviant than not barking. With these assumptions, we can show that $M \rightarrow S \rightarrow R \rightarrow E$ is a causal network. For we may assign *M*, *R*, and *E* the contrast 0, and *S* the contrast 1. Then: $E = 1$, rather than 0, depends upon $R = 1$, rather than 0; $R = 1$, rather than 0, depends upon $S = 3$, rather than 1; and $S = 3$, rather than 1, depends upon $M = 1$, rather than 0. Relative to this network, *S* is a departure variable and *E* its return, but both *S* and *E* have values which are more deviant than their contrasts. So the network is causal.

51. Of course, the right bell's ringing *was* a cause of Einstein's bark. So, like *Boulder*, *Short Circuit*, and figures 7 and 8, *Doorbells* provides a counterexample to the transitivity of causation. See Hall 2004 and Sartorio 2005. Cf. also Pearl 2000 (example 10.3.6) and Halpern and Pearl 2005.

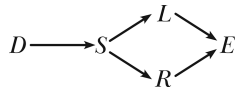
52. '$L := S = 2$' says that *L*'s value will be the truth-value of the proposition $S = 2$. That is: $L = 1$ if $S = 2$ and $L = 0$ if $S \neq 2$. Likewise for '$R := S = 3$'.

The assumptions that $S = 3$ is more deviant than $S = 1$ and $E = 1$ is more deviant than $E = 0$ aren't needed to show that there's a causal network from M to E. Even if they are not, the network



will be causal. For we may assign M, R, L, and E the contrast 0, and assign S the contrast 1 (note that L's contrast is the same as its value). Then: $E = 1$, rather than 0, depends upon $(L, R) = (0, 1)$, rather than $(0, 0)$; $R = 1$, rather than 0, depends upon $S = 3$, rather than 1; $L = 0$, rather than 0, depends upon $S = 3$, rather than 1; and $S = 3$, rather than 1, depends upon $M = 1$, rather than 0. In this network, there are no departures or returns, so the network is causal.

In contrast, so long as $S = 3$ is no more deviant than $S = 2$, there will be no causal network from D to E. We could assign D, R, and E the contrast 0, and assign S the contrast 2. Then: $E = 1$, rather than 0, depends upon $R = 1$, rather than 0; $R = 1$, rather than 0, depends upon $S = 3$, rather than 2; and $S = 3$, rather than 2, depends upon $D = 1$, rather than 0. But, relative to the network $D \rightarrow S \rightarrow R \rightarrow E$, S is a departure variable. Since its contrast is no more default than its value, this network is not causal. Nor is the network



causal. If D were to be 0, then S would be 2. And, if S were 2, then L would be 1 and R would be 0. So, if the path is to be causal, then (L, R) must be assigned the contrasts $(1, 0)$. But, if L were to be 1 and R were to be 0, then E would be 1. So E's contrast would not be distinct from its value. So the network is not causal.

The upshot is this: if Marty's pressing the button and Einstein's barking are both deviant, noninertial events, then the deviancy of Marty's pushing the button will be transferred to Einstein's barking, via a causal network. So Marty's pressing the button will be a productive cause of Einstein's barking. On the other hand, so long as the switch's directing a signal to the right is no more deviant than its directing a signal to the left, Doc's flipping the switch will not transfer any deviancy to Einstein's

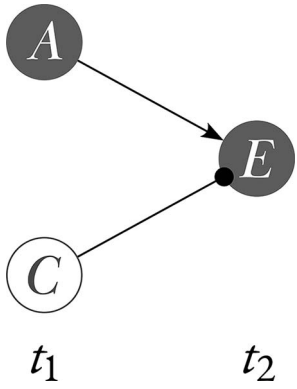


Figure 4. *Omission.*

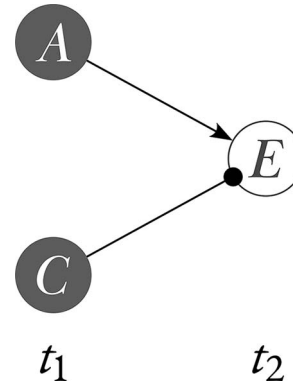


Figure 5. *Prevention.*

barking. Instead, Doc's flipping the switch merely *diverts* the deviancy of Marty's pushing the button to the right path. So Doc's flipping the switch will not be a productive cause of Einstein's barking.

If productive causation just is causation, then default, inertial states can be neither causes nor effects. Assuming that dormancy is the default state of a neuron, this means that *C*'s dormancy does not cause *E* to fire in the case of *Omission* from figure 4, nor does *C*'s firing cause *E* to not fire in the case of *Prevention* from figure 5 (both reproduced here).[53] If we find these consequences unacceptable, and we wish to insist that *Prevention* and *Omission* are both species of causation, then we may prefer the following theory of causation:

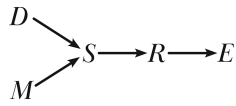
> Given a causal model **M** containing the variables in **C** and *E*, $\mathbf{C} = \mathbf{c}$ is a cause of $E = e$ in **M** iff, in **M**, there is a minimal causal network leading from **C** to *E*.

53. Both of these verdicts have defenders in the literature. Personally, I find the second verdict less intuitive than the first. I am currently inclined toward classifying *C*'s firing as a productive cause of *E*'s failure to fire in *Prevention* by appealing to a more nuanced account of when a variable value is *noninertial*. By way of explanation: I've some inclination to say that it would have been *inertial* for *E* to fire, given that *A* fired; and thus, that *E*'s failure to fire was a departure from that inertial behavior. (See note 12.) However, I won't explore this proposal any further here.

I call figure 4 a case of *omission* simply because *C*'s failure to fire is an omission, and *E*'s firing counterfactually depends upon this omission. I don't mean for the label to imply that this is an instance of causation *by* omission. Similarly, I call figure 5 a case of *prevention* merely because *E*'s failure to fire counterfactually depends upon *C*'s firing. I don't mean for the label 'prevention' to imply that this is an instance of causation, either.

Alternatively, we could allow **C**, but not *E*, to take on default values or more deviant contrasts. Or we could allow *E*, but not **C**, to take on a default value or a more deviant contrast. Because minimal causal networks are model-invariant (see proposition A.2 in the appendix), any of these accounts would be model-invariant. The kinds of values and contrasts we tolerate in our causes and effects is a free parameter of the theory.

I will say that a causal network, $\mathcal{N}$, from **C** to *E* is *minimal* iff there is no proper subnetwork of $\mathcal{N}$, leading from any subtuple of **C** to *E*, which is itself a causal network. In order for **C** to cause *E*, they must be connected by a *minimal* causal network. To understand why, return to the case of *Switch* from figure 11a. While there is no causal network leading from *D* to *E*, *there is* a causal network leading from the pair (D, M) to *E*:



Assign each of *D*, *M*, *S*, *R*, and *E* the contrast 0. Then, $E = 1$, rather than 0, depends upon $R = 1$, rather than 0; $R = 1$, rather than 0, depends upon $S = 3$, rather than 0; and $S = 3$, rather than 0, depends upon $(D, M) = (1, 1)$, rather than $(0, 0)$. In this network, *S* is a departure with return *E*, but both *S* and *E* have values more deviant than their contrasts. So this is a causal network. But *D* is not a joint cause of *E*'s firing, along with *M*. For $M \rightarrow S \rightarrow R \rightarrow E$ is a subnetwork of the causal network leading from (D, M) to *E*, and this subnetwork is causal. Requiring a causal network to be minimal prevents us from saying that *D*'s firing is a joint cause of *E*'s firing. More generally, it prevents us from counting as a joint cause any irrelevant factor 'free riding' on a causal network which it did nothing to help forge. In order to share in a causal network as a joint cause, you have to pull your weight.

Some theories impose a minimality condition on the variables in **C**. They say that **C** caused *E* only if no proper subtuple of **C** caused *E* (see, e.g., Halpern and Pearl 2001, 2005; Halpern 2016). These theories face difficulties with neuron systems like the one shown in figure 12. There, *C*'s firing is a joint cause of *E*'s firing. It, together with *A*, causes *E* to fire. However, if we were to impose a minimality condition on the variables in **C**, our theory would disagree. For even though there is a causal network from (A, C) to *E*, namely $A \rightarrow E \leftarrow C$, there is also a causal network from *A*

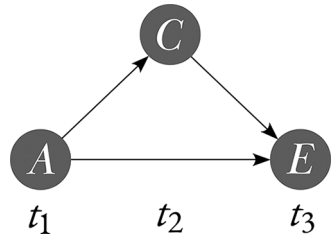


Figure 12.

alone to E, namely, $A \rightarrow C \rightarrow E \leftarrow A$.[54] Though the *tuple* (A, C) is not minimal, the *network* $A \rightarrow E \leftarrow C$ *is* minimal. So our theory tells us, correctly, that A and C jointly caused E to fire. (And also that A individually caused E to fire.)

Appendix. Technicalities

A notational convention: throughout this appendix, I will write things like '$\langle e, e^* \rangle$ locally depends upon $\langle \mathbf{c}, \mathbf{c}^* \rangle$' to mean that $E = e$, rather than e^*, locally depends upon $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$.

Proposition A.1. *If* $\mathbf{M} \models \mathbf{C} = \mathbf{c}^* \; \square\!\!\rightarrow \; E \neq e$, *then there is a causal network from some subtuple of* $\mathbf{C}$ *to* E *in* $\mathbf{M}$.

Proof. Let $\mathcal{N}$ be the union of every directed path leading from a member of $\mathbf{C}$ to E. We will show that if $\mathbf{M} \models \mathbf{C} = \mathbf{c}^* \; \square\!\!\rightarrow \; E \neq e$, then $\mathcal{N}$ is a causal network. (Since not every $C \in \mathbf{C}$ is guaranteed to be an ancestor of E, $\mathcal{N}$ may not be a causal network from $\mathbf{C}$ to E, but it will be a causal network from some subtuple of $\mathbf{C}$ to E.) Firstly, note that there are no departure or return variables on $\mathcal{N}$. For suppose there were a departure variable D with return R. Then, there would be a directed path from D to R, $D \rightarrow O_1 \rightarrow O_2 \rightarrow \ldots \rightarrow O_N \rightarrow R$, which is not included in $\mathcal{N}$. But there is a directed path from some member of $\mathbf{C}$ to D, and a directed path from R to E. So there is a directed path from some member of $\mathbf{C}$ to E which goes by way of the path $D \rightarrow O_1 \rightarrow O_2 \rightarrow \ldots \rightarrow O_N \rightarrow R$. Since $\mathcal{N}$ includes *every* directed path from $\mathbf{C}$ to E, this path must be included in $\mathcal{N}$. Contradiction. So there can be no departure or return variables on $\mathcal{N}$. For every variable V in the network $\mathcal{N}$, let 'v' be its actual value, and let its designated contrast, 'v^*', be the value it takes on in the counterfactual model $\mathbf{M}[\mathbf{C} \rightarrow \mathbf{c}^*]$. Since $\mathbf{M}[\mathbf{C} \rightarrow \mathbf{c}^*] \models E \neq e, e^* \neq e$, and E's contrast is distinct

54. I owe the objection to Ian Rosenberg and Clark Glymour (2018).

from its value. Now, take an arbitrary $D \notin \mathbf{C}$ which lies in $\mathcal{N}$. We now show that D's value, rather than its contrast, locally depends upon its $\mathcal{N}$-parents' values, rather than their contrasts. Let $\mathbf{P}_{\mathcal{N}}$ be the parents of D which lie in the network $\mathcal{N}$, and let $\mathbf{P}_{\overline{\mathcal{N}}}$ be the parents of D which do not lie in the network $\mathcal{N}$. By the construction of $\mathcal{N}$, $\mathbf{P}_{\overline{\mathcal{N}}}$ are not causal descendants of any member of $\mathbf{C}$. So, in the counterfactual model $\mathbf{M}[\mathbf{C} \rightarrow \mathbf{c}^*]$, $\mathbf{P}_{\overline{\mathcal{N}}}$ take on their actual values, $\mathbf{p}_{\overline{\mathcal{N}}}$. Since $\mathbf{M}[\mathbf{C} \rightarrow \mathbf{c}^*] \models D = d^*$,

$$\phi_D(\mathbf{p}^*_{\mathcal{N}}, \mathbf{p}_{\overline{\mathcal{N}}}) = d^*$$

So $\langle d, d^* \rangle$ locally depends upon $\langle \mathbf{p}_{\mathcal{N}}, \mathbf{p}^*_{\mathcal{N}} \rangle$. D was arbitrary, so the same goes for every variable in the network $\mathcal{N}$, except for those in $\mathbf{C}$. So there is a causal network running from (some subtuple of) $\mathbf{C}$ to E. □

Remark. The proposition shows us that counterfactual dependence suffices for a causal network, but this causal network need not be *minimal*. If $\mathbf{C}$ is a singleton, however, then counterfactual dependence will suffice for a minimal causal network. For counterfactual dependence of $E = e$ on $\mathbf{C} = c$ means that there is some causal network from $\mathbf{C}$ to E. Perhaps this network is not minimal, but no matter—if it is not minimal, then some subnetwork of it will be both causal and minimal. So there will be some minimal causal network from $\mathbf{C}$ to E.

Lemma A.1. *Given a causal model* $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, *with* $U \in \mathbf{U}$, $\mathbf{C} \subset \mathbf{U} \cup \mathbf{V}$, $E \in \mathbf{V}$, *and* $U \notin \mathbf{C}$, $\mathcal{N}$ *is a causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}$ *iff* $\mathcal{N}$ *is a causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}^{-U}$.

Proof. Suppose that $\mathcal{N}$ is a causal network from $\mathbf{C}$ to E in $\mathbf{M}$. The exogenous $U \in \mathbf{U}$ will not be in this network, so removing it will not affect any of the local dependence relationships between any of the variables in $\mathcal{N}$. Nor will it affect whether any departure or return variables along $\mathcal{N}$ have values more deviant than their contrasts. So $\mathcal{N}$ will be a causal network from $\mathbf{C}$ to E in $\mathbf{M}^{-U}$. Suppose, on the other hand, that $\mathcal{N}$ was not a causal network from $\mathbf{C}$ to E in $\mathbf{M}$. If $\mathcal{N}$ is a network from $\mathbf{C}$ to E, then the exogenous $U \in \mathbf{U}$ is not on this network, and removing it will not affect the local dependence relationships between any of the variables on $\mathcal{N}$, nor whether any departure and return variables have values more deviant than their contrasts. So removing U will not make $\mathcal{N}$ into a causal network from $\mathbf{C}$ to E. So $\mathcal{N}$ will not be a causal network from $\mathbf{C}$ to E in $\mathbf{M}^{-U}$. □

Definition A.1. If V is an interpolated variable in $\mathbf{M}$ with parent Pa and child Ch, and $\mathcal{N}$ is a network in $\mathbf{M}$ (which may or may not contain the directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$), then let $\mathcal{N} - V$ be the network in $\mathbf{M}^{-V}$ defined as follows: if V lies along $\mathcal{N}$, then $\mathcal{N} - V$ is $\mathcal{N}$, minus the directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$, and plus the new directed edge $Pa \rightarrow Ch$; and if V does not lie along $\mathcal{N}$, then $\mathcal{N} - V$ is just $\mathcal{N}$.

Definition A.2. If V is an interpolated variable in $\mathbf{M}$ with parent Pa and child Ch, and $\mathcal{N}$ is a network in $\mathbf{M}^{-V}$ (which may or may not contain the directed edge $Pa \rightarrow Ch$), then let $\mathcal{N} + V$ be the network in $\mathbf{M}$ defined as follows: if $\mathcal{N}$ includes $Pa \rightarrow Ch$, then $\mathcal{N} + V$ is $\mathcal{N}$, minus $Pa \rightarrow Ch$, and plus the directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$; and if $\mathcal{N}$ does not include $Pa \rightarrow Ch$, then $\mathcal{N} + V$ is just $\mathcal{N}$.

Lemma A.2. *Given a causal model* $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, *with* $V \in \mathbf{V}$, $\mathbf{C} \subset \mathbf{U} \cup \mathbf{V}$, $E \in \mathbf{V}$, *and* $V \notin \mathbf{C} \cup (E)$, *if* V *is inessential in* $\mathbf{M}$, *then: (a) if* $\mathcal{N}$ *is a causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}$, *then* $\mathcal{N} - V$ *is a causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}^{-V}$; *and (b) if* $\mathcal{N}$ *is a causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}^{-V}$, *then* $\mathcal{N} + V$ *is a causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}$.

Proof. Start with part (a). Suppose that $\mathcal{N}$ is a causal network from $\mathbf{C}$ to E in $\mathbf{M}$. Since V is inessential, it has a single parent, Pa, and a single child, Ch (and Pa is not a parent of Ch). Let their actual values in $\mathbf{M}$ be v, pa, and ch, respectively. There are two possibilities: either (A) V does not lie on $\mathcal{N}$, or (B) it does. In case (A), removing V may introduce new local dependence relationships between Pa and Ch, but it will not alter any local dependence relations between any of the variables on $\mathcal{N}$ and their $\mathcal{N}$-parents. Since, in $\mathbf{M}$, each variable in $\mathcal{N}$, rather than its contrast, locally depends upon its $\mathcal{N}$-parents' values, rather than their contrasts, in $\mathbf{M}^{-V}$, each variable in $\mathcal{N} - V = \mathcal{N}$, rather than its contrast, will still locally depend upon its $\mathcal{N} - V$-parents' values, rather than their contrasts. For any departure or return variables in $\mathcal{N}$, removing V will not affect whether these variables are departure/return variables, nor whether their values are more deviant than their contrasts. So, in case (A), $\mathcal{N} - V$ will be a causal network in $\mathbf{M}^{-V}$. In case (B), V lies on $\mathcal{N}$. Then, Pa and Ch must lie on $\mathcal{N}$ as well. Let $\mathbf{R}_{\mathcal{N}}$ be Ch's causal parents other than V that lie in the network $\mathcal{N}$ (if such there be); let their actual values be $\mathbf{r}_{\mathcal{N}}$ and their designated contrasts, $\mathbf{r}^*_{\mathcal{N}}$. Similarly, let $\mathbf{R}_{\overline{\mathcal{N}}}$ be Ch's causal parents that don't lie on the network $\mathcal{N}$ (if such there be), and let their actual values be $\mathbf{r}_{\overline{\mathcal{N}}}$. Then, in $\mathbf{M}$, there are some v^*, pa^*, and ch^* such that $\langle ch, ch^* \rangle$

locally depends upon $\langle \mathbf{r}_{\mathcal{N}} \cup (v), \mathbf{r}^*_{\mathcal{N}} \cup (v^*)\rangle$ and $\langle v, v^*\rangle$ locally depends upon $\langle pa, pa^*\rangle$. Since *Pa* is *V*'s only parent, we can conclude that

(1) $\phi_V(pa^*) = v^*$

And since $\langle ch, ch^*\rangle$ locally depends upon $\langle \mathbf{r}_{\mathcal{N}} \cup (v), \mathbf{r}^*_{\mathcal{N}} \cup (v^*)\rangle$, we can conclude that

(2) $\phi_{Ch}(v^*, \mathbf{r}^*_{\mathcal{N}}, \mathbf{r}_{\overline{\mathcal{N}}}) = ch^*$

By the construction of $\mathbf{M}^{-V}$, it contains the structural equation

$$Ch := \phi_{Ch}(\phi_V(Pa), \mathbf{R}_{\mathcal{N}}, \mathbf{R}_{\overline{\mathcal{N}}})$$

Note that from (1) and (2), it follows that

$$\phi_{Ch}(\phi_V(pa^*), \mathbf{r}^*_{\mathcal{N}}, \mathbf{r}_{\overline{\mathcal{N}}}) = ch$$

So, in $\mathbf{M}^{-V}$, $\langle ch, ch^*\rangle$ locally depends upon $\langle \mathbf{r}_{\mathcal{N}} \cup (pa), \mathbf{r}^*_{\mathcal{N}} \cup (pa^*)\rangle$. Removing *V* will not affect whether any variables are departure or return variables, relative to $\mathcal{N}$, nor whether departure and return variables have values more deviant than their contrasts. So $\mathcal{N} - V$ will be a causal network in $\mathbf{M}^{-V}$.

To establish part (b), suppose that $\mathcal{N}$ is a causal network from $\mathbf{C}$ to *E* in $\mathbf{M}^{-V}$. $\mathcal{N}$ either (A) includes the directed edge $Pa \rightarrow Ch$, or (B) it doesn't. If (A), then there must be some pa^*, ch^*, and $\mathbf{r}^*_{\mathcal{N}}$ such that $\langle ch, ch^*\rangle$ locally depends upon $\langle \mathbf{r}_{\mathcal{N}} \cup (pa), \mathbf{r}^*_{\mathcal{N}} \cup (pa^*)\rangle$. ($\mathbf{R}_{\mathcal{N}}$ are *Ch*'s $\mathcal{N}$-parents other than *Pa*, if such there be.) So

(3) $\phi_{Ch}(\phi_V(pa^*), \mathbf{r}^*_{\mathcal{N}}, \mathbf{r}_{\overline{\mathcal{N}}}) = ch^*$

($\mathbf{R}_{\overline{\mathcal{N}}}$ are the parents of *Ch* which do not lie on the network $\mathcal{N}$, if such there be.) Let v^* be the value of *V* such that $v^* = \phi_V(pa^*)$. Then, it follows from (3) that $\langle ch, ch^*\rangle$ will locally depend upon $\langle \mathbf{r}_{\mathcal{N}} \cup (v), \mathbf{r}^*_{\mathcal{N}} \cup (v^*)\rangle$ in $\mathbf{M}$. Including *V* will not affect which variables are departure/return variables, nor whether their values are more deviant than their contrasts. So $\mathcal{N} + V$ will be a causal network from $\mathbf{C}$ to *E* in $\mathbf{M}$. If (B), then $\mathcal{N} + V = \mathcal{N}$ will also be a causal network from $\mathbf{C}$ to *E* in $\mathbf{M}$, since including the interpolated variable *V* will not alter any of the local dependence relationships among any of the variables other than *Pa* and *Ch*, nor will it affect which variables are departure/returns relative to $\mathcal{N}$, nor whether their values are more deviant than their contrasts. □

Proposition A.2. *Minimal causal networks are model-invariant. That is*: *(a) given a causal model* $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, *with* $U \in \mathbf{U}$, $\mathbf{C} \subset \mathbf{U} \cup \mathbf{V}$, $E \in \mathbf{V}$, *and* $U \notin \mathbf{C}$, *there is a minimal causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}$ *iff there is a minimal causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}^{-U}$, *and (b) given a causal model* $\mathbf{M} = (\mathbf{U}, \mathbf{u}, \mathbf{V}, \mathbf{E}, \geq)$, *with* $V \in \mathbf{V}$, $\mathbf{C} \subset \mathbf{U} \cup \mathbf{V}$, $E \in \mathbf{V}$, *and* $V \notin \mathbf{C} \cup (E)$, *if* V *is inessential in* $\mathbf{M}$, *then there is a minimal causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}$ *iff there is a minimal causal network from* $\mathbf{C}$ *to* E *in* $\mathbf{M}^{-V}$.

Proof. Begin with part (b): suppose there is a minimal causal network from $\mathbf{C}$ to E in $\mathbf{M}$. Then, there is a causal network, $\mathcal{N}$, from $\mathbf{C}$ to E in $\mathbf{M}$, and there is no proper subnetwork of $\mathcal{N}$, from any subtuple of $\mathbf{C}$ to E in $\mathbf{M}$. By lemma A.2, $\mathcal{N} - V$ is a causal network in $\mathbf{M}^{-V}$. Suppose (for reductio) that this causal network is not minimal. Then, there is some proper subnetwork of $\mathcal{N} - V$, $\mathcal{N}^*$, from some subtuple of $\mathbf{C}$ to E in $\mathbf{M}^{-V}$ which is causal. By lemma A.2, $\mathcal{N}^* + V$ is a causal network in $\mathbf{M}$. If $\mathcal{N}^*$ is a proper subnetwork of $\mathcal{N} - V$ in $\mathbf{M}^{-V}$, then $\mathcal{N}^* + V$ is a proper subnetwork of $\mathcal{N}$ in $\mathbf{M}$. So in $\mathbf{M}$ there is a proper subnetwork of $\mathcal{N}$, from some subtuple of $\mathbf{C}$ to E, which is causal. So $\mathcal{N}$ is not a *minimal* causal network in $\mathbf{M}$. Contradiction. So $\mathcal{N} - V$ is a minimal causal network in $\mathbf{M}^{-V}$.

Going in the other direction, suppose that there is a minimal causal network from $\mathbf{C}$ to E in $\mathbf{M}^{-V}$. So there is a causal network, $\mathcal{N}$, in $\mathbf{M}^{-V}$, and there is no proper subnetwork of $\mathcal{N}$, from any subtuple of $\mathbf{C}$ to E in $\mathbf{M}^{-V}$. By lemma A.2, $\mathcal{N} + V$ is a causal network in $\mathbf{M}$. Suppose (for reductio) that this causal network is not minimal. Then, there is some proper subnetwork of $\mathcal{N}$, $\mathcal{N}^*$, from some subtuple of $\mathbf{C}$ to E in $\mathbf{M}$, which is causal. By lemma A.2, $\mathcal{N}^* - V$ is a causal network from some subtuple of $\mathbf{C}$ to E in $\mathbf{M}^{-V}$. If $\mathcal{N}^*$ is a proper subnetwork of $\mathcal{N} + V$ in $\mathbf{M}$, then $\mathcal{N}^* - V$ is a proper subnetwork of $\mathcal{N}$ in $\mathbf{M}^{-V}$. So in $\mathbf{M}^{-V}$ there is a proper subnetwork of $\mathcal{N}$, from some subtuple of $\mathbf{C}$ to E, which is causal. So $\mathcal{N}$ is not a minimal causal network in $\mathbf{M}^{-V}$. Contradiction. So $\mathcal{N} + V$ is a *minimal* causal network from $\mathbf{C}$ to E in $\mathbf{M}$.

The proof of part (a) is exactly the same, with lemma A.2 swapped out for lemma A.1, $\mathbf{M}^{-V}$ swapped out for $\mathbf{M}^{-U}$, $\mathcal{N} - V$ and $\mathcal{N} + V$ swapped out for $\mathcal{N}$, and $\mathcal{N}^* - V$ and $\mathcal{N}^* + V$ swapped out for $\mathcal{N}^*$. □

References

Andreas, Holger, and Mario Günther. 2018. "A Ramsey Test Analysis of Causation for Causal Models." *British Journal for the Philosophy of Science*. Published ahead of print, December 10. doi.org/10.1093/bjps/axy074.

Andreas, Holger, and Mario Günther. 2020. "Causation in Terms of Production." *Philosophical Studies* 177, no. 6: 1565–91. doi.org/10.1007/s11098-019-01275-3.

Armstrong, David. 2004. "Going through the Open Door Again: Counterfactuals vs. Singularist Theories of Causation." In Collins, Hall, and Paul 2004, 445–57.

Beckers, Sander, and Joost Vennekens. 2017. "The Transitivity and Asymmetry of Actual Causation." *Ergo* 4, no. 1: 1–27.

Beckers, Sander, and Joost Vennekens. 2018. "A Principled Approach to Defining Actual Causation." *Synthese* 195, no. 2: 835–62.

Briggs, R.A. 2012. "Interventionist Counterfactuals." *Philosophical Studies* 160, no. 1: 139–66.

Collins, J., Ned Hall, and L. A. Paul, eds. 2004. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Dowe, Phil. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Fair, David. 1979. "Causation and the Flow of Energy." *Erkenntnis*14, no. 3: 219–50.

Galles, David, and Judea Pearl. 1998. "An Axiomatic Characterization of Causal Counterfactuals." *Foundations of Science* 3, no. 1: 151–82.

Gallow, J. Dmitri. 2016. "A Theory of Structural Determination." *Philosophical Studies* 173, no. 1: 159–86.

Gallow, J. Dmitri. n.d. "Model-Variance in Theories of Token Causation." Unpublished manuscript.

Hall, Ned. 2004. "Two Concepts of Causation." In Collins, Hall, and Paul 2004: 225–76.

Hall, Ned. 2007. "Structural Equations and Causation." *Philosophical Studies* 132, no. 1: 109–36.

Halpern, Joseph Y. 2008. "Defaults and Normality in Causal Structures." In *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, edited by Gerhard Brewka and Jérôme Lang, 198–208. Menlo Park, CA: AAAI Press.

Halpern, Joseph Y. 2016. *Actual Causality*. Cambridge, MA: MIT Press.

Halpern, Joseph Y., and Judea Pearl. 2001. "Causes and Explanations: A Structural-Model Approach." Pt. 1, "Causes." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, edited by John Breese and Daphne Koller, 194–202. San Francisco: Morgan Kaufman.

Halpern, Joseph Y., and Judea Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." *British Journal for the Philosophy of Science* 56, no. 4: 843–87.

Halpern, Joseph Y., and Christopher Hitchcock. 2015. "Graded Causation and Defaults." *British Journal for the Philosophy of Science* 66, no. 2: 413–57.

Hitchcock, Christopher. 1996a. "The Role of Contrast in Causal and Explanatory Claims." *Synthese* 107, no. 3: 395–419.

Hitchcock, Christopher. 1996b. "Farewell to Binary Causation." *Canadian Journal of Philosophy* 26, no. 2: 267–82.
Hitchcock, Christopher. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98, no. 6: 273–99.
Hitchcock, Christopher. 2007a. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review* 116, no. 4: 495–532.
Hitchcock, Christopher. 2007b. "What's Wrong with Neuron Diagrams?" In *Causation and Explanation*, edited by Joseph Keim Campbell, Michael O'Rourke, and Harry Silverstein, 69–92. Cambridge, MA: MIT Press.
Hitchcock, Christopher. 2011. "Trumping and Contrastive Causation." *Synthese* 181, no. 2: 227–40.
Hitchcock, Christopher, and Joshua Knobe. 2009. "Cause and Norm." *Journal of Philosophy* 106, no. 11: 587–612.
Huber, Franz. 2013. "Structural Equations and Beyond." *Review of Symbolic Logic* 6, no. 4: 709–32.
Kahneman, Daniel, and Dale T. Miller. 1986. "Norm Theory: Comparing Reality to Its Alternatives." *Psychological Review* 94, no. 2: 136–53.
Lewis, David K. 1973. "Causation." *Journal of Philosophy* 70, no. 17: 556–67.
Lewis, David K. 1986. "Postscripts to 'Causation'." In *Philosophical Papers*, vol. 2: 172–213. Oxford: Oxford University Press.
Lewis, David K. 2004. "Causation as Influence." In Collins, Hall, and Paul 2004: 75–106.
Livengood, Jonathan. 2013. "Actual Causation in Simple Voting Scenarios." *Noûs* 47, no. 2: 316–45.
Mackie, John L. 1965. "Causes and Conditions." *American Philosophical Quarterly* 2, no. 4: 245–55.
Maslen, Cei. 2004. "Causes, Contrasts, and the Nontransitivity of Causation." In Collins, Hall, and Paul 2004: 341–57.
Maudlin, Tim. 2004. "Causation, Counterfactuals, and the Third Factor." In Collins, Hall, and Paul 2004: 419–43.
McDermott, Michael. 1995. "Redundant Causation." *British Journal for the Philosophy of Science* 46, no. 4: 523–44.
McGrath, Sarah. 2005. "Causation by Omission: A Dilemma." *Philosophical Studies* 123, no. 1–2: 125–48.
Menzies, Peter. 2004. "Causal Models, Token Causation, and Processes." *Philosophy of Science* 71, no. 5: 820–32.
Menzies, Peter. 2006. "A Structural Equations Account of Negative Causation." http://philsci-archive.pitt.edu/2962 (deposited October 10, 2006).
Paul, L. A. 2004. "Aspect Causation." In Collins, Hall, and Paul 2004: 205–24.
Paul, L. A., and Ned Hall. 2013.*Causation: A User's Guide*. Oxford: Oxford University Press.
Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Rosenberg, Ian, and Clark Glymour. 2018. "Review of *Actual Causality*, by Joseph Y. Halpern." *British Journal for the Philosophy of Science Review of Books*. www.thebsps.org/2018/07/joseph-y-halpern-actual-causality/.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Salmon, Wesley. 1994. "Causality without Counterfactuals." *Philosophy of Science* 61, no. 2: 297–312.

Sartorio, Carolina. 2005. "Causes as Difference-Makers." *Philosophical Studies* 123, no. 1–2: 71–98.

Sartorio, Carolina. 2016. *Causation and Free Will*. Oxford: Oxford University Press.

Schaffer, Jonathan. 2000. "Causation by Disconnection." *Philosophy of Science* 67, no. 2: 285–300.

Schaffer, Jonathan. 2003. "Overdetermining Causes." *Philosophical Studies* 114, no. 1–2: 23–45.

Schaffer, Jonathan. 2004. "Trumping Preemption." In Collins, Hall, and Paul 2004: 59–75.

Schaffer, Jonathan. 2005. "Contrastive Causation." *Philosophical Review* 114, no. 3: 297–328.

Schaffer, Jonathan. 2012a. "Causal Contextualism." In *Contrastivism in Philosophy*, edited by Martijn Blaauw, 35–63. New York: Routledge.

Schaffer, Jonathan. 2012b. "Disconnection and Responsibility." *Legal Theory* 18, no. 4: 399–435.

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research* 66, no. 1: 81–103.

Weslake, Brad. Forthcoming. "A Partial Theory of Actual Causation." *British Journal for the Philosophy of Science*.

Wolff, J. E. 2016. "Using Defaults to Understand Token Causation." *Journal of Philosophy* 113, no. 1: 5–26.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Yablo, Stephen. 2002. "De Facto Dependence." *Journal of Philosophy* 99, no. 3: 130–48.

Yablo, Stephen. 2004. "Advertisement for a Sketch of an Outline of a Prototheory of Causation." In Collins, Hall, and Paul 2004: 119–38.