# Causation, Production, and Dependence

J. Dmitri Gallow

Abstract

I provide a theory of causation formulated within the causal modeling framework. In contrast to its predecessors, this theory is model-invariant in the following sense: if the theory says that $C$ caused (didn't cause) $E$ in a causal model $\mathcal{M}$, then it will continue to say that $C$ caused (didn't cause) $E$ once we've removed an inessential variable from $\mathcal{M}$. On this theory, we can understand causation as a model-invariant generalization of a relation of causal production. Begin by saying that $C$ produces $E$ iff they are linked by an uninterrupted process propagating non-inertial states. Given this definition, whether we say that $C$ produces $E$ will depend upon whether we include or remove inessential variables lying along, or feeding into, the path from $C$ to $E$. Weakening production so as to make the relation model-invariant delivers causation.

Causal models formally represent the networks of determination in which causes and effects are ensconced. 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 will also tell us whether one variable is a token cause of another.[1] To this end, a number of authors have developed theories of causation within the causal modelling framework.[2] Lots of good work has been done on this front, but the theories developed to date have an awkward consequence: adding or removing

Draft of October 14, 2018. Word count: 16,012
Comments appreciated. ✉: jdmitrigallow@pitt.edu

[1] Token causation is sometimes called 'singular causation' or 'actual causation'. Token causal relations are the causal relations described with token causal claims—sentences of the form '*c*'s *F*-ing caused *e* to *G*', where *c*'s *F*-ing and *e*'s *G*-ing are token events (*e.g.*, 'Chris's drinking caused him to contract 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 represented in a causal model. (Looking ahead to §1, in figure 1, whether $B$ fires causally determines whether $E$ does, but $B$'s failure to fire is not a token cause of $E$'s firing.) Because my focus will be on token causation throughout, 'cause' should always be understood to mean 'token cause'.

[2] See, *e.g.*, Halpern & Pearl (2001, 2005), Menzies (2004, 2006), Hitchcock (2001, 2007), Woodward (2003), Hall (2007), Halpern (2008, 2016), and Weslake (forthcoming).

an inessential variable from a model will lead these theories to revise their verdicts about whether two variables are causally related. Attend to an additional, inessential, variable interpolated along the path leading from $C$ to $E$, and these theories will change their mind about whether $C$ caused $E$. Attend to an additional, inessential, variable feeding into the path from $C$ to $E$, and these theories will likewise change their mind about whether $C$ caused $E$.[3,4]

Below, I will present a model-invariant a theory of causation. This theory's verdicts about whether $C$ caused $E$ will not change when inessential variables along or leading into the path from $C$ to $E$ are removed. I will develop this theory in stages by considering some standard problem cases from the literature—preemptive overdetermination (§3), counterexamples to transitivity (§4), preemptive prevention (§5), and, finally, symmetric overdetermination (§7). As the account is developed, revisions and emendments will be motivated by two constraints: firstly, that the theory agree with some widely shared causal judgments, like the judgment that a preemptive overdeterminer is a cause and the judgment that a 'switch' is not; and secondly, that the theory be model-invariant. One of my goals below will be to persuade you that these two constraints leave very few choice points, and that the theory I end up presenting is the natural destination towards which they lead.

This theory allows us to understand causation as a model-invariant generalization of a relation of causal production. Begin by understanding a productive process as a sequence of variables, each of which takes on some deviant or non-inertial value. Say that $C$ produces $E$ iff they are linked by such a productive process. This relation of production well characterizes our most paradigmatic and uncontroversial causal judgments, but it is a model-variant relation. The theory I develop below says that causation is weakened production; production is sufficient, but not necessary, for causation. Every bit of this weakening is required in order for causation to be a model-invariant relation (§6).

## 1 Causal Models

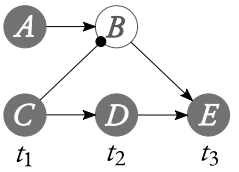
As I'll be using the term here, a *causal model* consists of 5 components: a collection of exogenous variables, $\mathcal{U}$, an assignment of values to those variables, $\mathbf{u}$, a collection of endogenous variables, $\mathcal{V}$, a system of *structural equations*, one for each endogenous variable in $\mathcal{V}$, and a specification of which variable values are *default* and which are *deviant*.[5]

---

[3] I show this in a companion paper.

[4] I'll get more precise about the term 'inessential' in §2 below.

[5] A word on notation: variables will be denoted with uppercase italic letters ($A, B, C, \ldots$), while their values will be denoted with the corresponding lowercase italic letters ($a, b, c, \ldots$). Vectors will be indicated with boldface, $\mathbf{V}$, or calligraphic letters, $\mathcal{V}$. I will use uppercase for a vector of




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

FIGURE 1: On the left, a neuron network of a case of *Preemptive Overdetermination*. On the right, the *canonical causal model*, $\mathcal{M}_1$, of this neuron network. (For all variables, the value 0 is default, and the value 1 is deviant.)

CAUSAL MODELS

A causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$ is a 5-tuple of

(a) A vector, $\mathcal{U} = (U_1, U_2, \ldots, U_M)$, of *exogenous* variables;

(b) An assignment of values, $\mathbf{u} = (u_1, u_2, \ldots, u_M)$, to $\mathcal{U}$;

(c) A vector $\mathcal{V} = (V_1, V_2, \ldots, V_N)$, of *endogenous* variables;

(d) A vector $\mathcal{E} = (\phi_{V_1}, \phi_{V_2}, \ldots, \phi_{V_N})$ of *structural equations*, one for each endogenous variable $V_i \in \mathcal{V}$; and

(e) A specification, $\mathcal{D}$, of which values of the variables in $\mathcal{U} \cup \mathcal{V}$ are *default*, *normal*, or *inertial*, and which are *deviations* therefrom.

To see how a causal model represents structures of causal determination, consider the LEWISIAN 'neuron network' shown in figure 1. Here's how to read the diagram in figure 1: for every time listed at the bottom, the neurons above it can either fire or not fire at that time. If a 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 as representational tools. This is not how I will be understanding them here. Rather, I will be understanding the neuron networks displayed in these diagrams as the reality to be represented with a causal model (to emphasize this, I will refer to

variables and lowercase for a vectors of values. The Greek letter $\phi$, subscripted with a variable, will stand for a function, and I will often use just $\ulcorner \phi_V \urcorner$ to stand for an entire structural equation like $V \coloneqq \phi_V(X, Y, Z)$. Throughout, I will apply set-theoretic notation to *vectors* of variables. Thus, $\mathcal{U} \cup \mathcal{V}$ is a vector containing all and only the variables in either $\mathcal{U}$ or $\mathcal{V}$, $\mathcal{V} \setminus \mathbf{X}$ is a vector containing all and only the variables in $\mathcal{V}$, except for those in $\mathbf{X}$, and so on. There will in general be many such vectors, depending upon an arbitrary choice of order. It won't matter which of these an expression like '$\mathcal{U} \cup \mathcal{V}$' denotes.

them as 'neuron networks', rather than the more common 'neuron diagrams'). To represent the neuron network 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.[6] Both $A$ and $C$ are *exogenous* variables—variables whose values are not causally determined by the values of the other variables in the model. And, since both of those neurons fire at $t_1$, the exogenous assignment will tell us that $A = C = 1$. $B$, $D$, and $E$ will be *endogenous* variables—variables whose values are causally determined by the values of the other variables in the model. The structural equations in $\mathcal{E}$ 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 iff either $B$ or $D$ does. 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 the model $\mathcal{M}_1$, the structural equation $B := A \wedge \neg C$, together with the exogenous assignment $A = C = 1$, tells us that $B = 0$.[7] 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 $\mathcal{E}$ encode information about the direction of causal determination, we cannot re-arrange $D := C$ 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, $\phi_V$, 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 $V$. It is for this reason that I use ':=', rather than the symmetric '=', in structural equations.

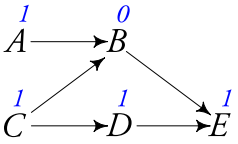
Given a causal model, $\mathcal{M}$, we may construct a *causal graph* which displays the causal determination structure between the variables in $\mathcal{U} \cup \mathcal{V}$, as follows: if a variable $U$ appears on the right-hand-side of $V$'s structural equation $\phi_V$, then place a directed edge between $U$ and $V$, with its tail at $U$ and its head at $V$, $U \to V$. 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

---

[6] Thus, I am using '$A$' to stand for *both* the neuron *and* the variable which represents whether $A$ fires at $t_1$. Context will disambiguate.

[7] $x \wedge y$, $x \vee y$, and $\neg x$ are the functions $\min\{x, y\}$, $\max\{x, y\}$, and $1 - x$, respectively.




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 directly causally determined, the graph on its own does not tell us precisely *how* the values of the endogenous variables are causally determined. For that information, we must look to the structural equations in $\mathcal{E}$.

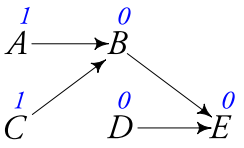
It is common to use the metaphor of genealogy to describe the structural relations between variables displayed in a graph. For instance, $B$ and $D$ are $E$'s causal parents, and $C$'s causal children. Similarly, $B$, $D$, and $E$ are $C$'s causal descendants. (Throughout, I will assume that no variable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

Finally, our causal model should specify, for each variable, which values of that variable are *default*, *inertial*, or *normal* values, and which values are *deviant*, *non-inertial*, *departures* from normality. In the case of our neuron network 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 an abnormal deviation from that default, inertial state. I will assume likewise for every other neuron network in this paper. (The reader may be curious why this kind of information is included in a causal model—I will come to that shortly, in §1.2 below.)

To construct the causal model $\mathcal{M}_1$ from the neuron network in figure 1, 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 at the far left-hand-side of the diagram were made exogenous, and assigned the values corresponding to the actual state of their neurons. We then wrote down equations describing how the state of each endogenous neuron is directly causally determined by the other neurons in the network. Let's call the causal model that we construct in this way from a given neuron network the *canonical causal model* of that neuron network. I'll assume throughout that the canonical causal model of a given neuron network is correct.

### 1.1 Counterfactual Causal Models

Given a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with some vector of variables $\mathbf{A} \subseteq \mathcal{U} \cup \mathcal{V}$, we may construct a *counterfactual* model, in which the variables in $\mathbf{A}$ have been intervened upon so as to hold their values fixed at $\mathbf{a}$, as follows: We *remove* any endogenous variables in $\mathbf{A}$ from the endogenous variable vector $\mathcal{V}$, and




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

FIGURE 2: On the right, the counterfactual model $\mathcal{M}_1[D \rightarrow 0]$ (for all variables, 0 is default, and 1 is deviant). On the left, its associated causal graph.

add them to the exogenous variable vector, $\mathcal{U}$. Next, we remove the structural equations of any endogenous variables in $\mathbf{A}$ from the structural equations vector $\mathcal{E}$, and change the exogenous assignment vector $\mathbf{u}$ so that it assigns the values in $\mathbf{a}$ to the variables in $\mathbf{V}$. The specification of which variable values are default and which are deviant will remain unchanged.

COUNTERFACTUAL CAUSAL MODEL
Given a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, including the variables $\mathbf{A}$, and given the assignment of values $\mathbf{a}$ to $\mathbf{V}$, the counterfactual model

$$\mathcal{M}[\mathbf{A} \rightarrow \mathbf{a}] = (\mathcal{U}^*, \mathbf{u}^*, \mathcal{V}^*, \mathcal{E}^*, \mathcal{D}^*)$$

is the model such that:

(a) $\mathcal{U}^* = \mathcal{U} \cup \mathbf{A}$
(b) $\mathbf{u}^* = \mathbf{u} + \mathbf{a}$
(c) $\mathcal{V}^* = \mathcal{V} \setminus \mathbf{A}$
(d) $\mathcal{E}^* = \mathcal{E} \setminus (\phi_{A_i} \mid A_i \in \mathbf{A})$
(e) $\mathcal{D}^* = \mathcal{D}$

For instance, figure 2 displays the counterfactual model $\mathcal{M}_1[D \rightarrow 0]$ in which we have intervened so as 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 $C$. Rather, $D$ has been 'exogenized', and it has been given the exogenous assignment 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 counterfactuals.[8] According to this semantics, a counterfactual "Had $\mathbf{A}$ taken on the values $\mathbf{a}$, then it would have been that C" (where C is any Boolean combination of variable values) is true in a causal model $\mathcal{M}$ just in case C is true in the *counterfactual model* in which you've intervened so as to set $\mathbf{A}$ to the values $\mathbf{a}$, $\mathcal{M}[\mathbf{A} \rightarrow \mathbf{a}]$.

CAUSAL COUNTERFACTUALS
If C is a proposition about the values of the variables in a causal

[8] For more on this semantics, see GALLES & PEARL (1998), BRIGGS (2012), and HUBER (2013).

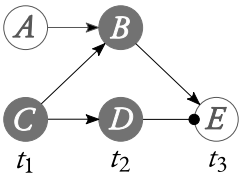



FIGURE 3: *Short Circuit*

> model $\mathcal{M}$, and $\mathcal{M}$ contains the variables in $\mathbf{A}$, then the causal counterfactual $\mathbf{A} = \mathbf{a} \boxdot\!\!\rightarrow \mathrm{C}$ is true in $\mathcal{M}$ iff C is true in the counterfactual model $\mathcal{M}[\mathbf{A} \rightarrow \mathbf{a}]$,
>
> $$\mathcal{M} \models \mathbf{A} = \mathbf{a} \boxdot\!\!\rightarrow \mathrm{C} \iff \mathcal{M}[\mathbf{A} \rightarrow \mathbf{a}] \models \mathrm{C}$$

## 1.2 Defaults and Deviancy

The neuron network 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 $C$ runs to completion, the causal process initiated with $A$ 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. For a case with the same causal structure, consider TAX CUT.

> 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 the 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. Still, 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. Sammy, on the other hand, is a genuine cause of the proposal's passing.

Consider the neuron network shown in figure 3. 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

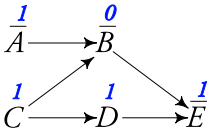
same time, neutralizes that very threat. I follow Hall (2007) in calling this neuron network a 'short circuit'.[9] For a case with the same causal structure, consider Boulder,[10]

> 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.

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. The boulder's becoming dislodged did not cause Matthew to survive. Nor did $C$'s firing cause $E$ to remain dormant in the neuron network from figure 3.

As Hall (2007) notes, we may write down a system of structural equations for *Short Circuit* which is isomorphic to the canonical causal 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* fire. Similarly, let $\overline{B}$ and $\overline{E}$ be variables which take on the value 1 if their associated neurons *don't* fire, and take on the value 0 if they *do* fire. 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 amongst these variables.

$$\begin{aligned} \overline{E} &:= \overline{B} \vee D \\ D &:= C \\ \overline{B} &:= \overline{A} \wedge \neg C \end{aligned}$$




$E$ *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 with 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.

---

[9] See also Lewis (2004, p. 97–99), in which the same structure is called an *inert network*.

[10] The case is attributed to an early draft of Hall (2004) by Hitchcock (2001).

It is natural to think of the non-firing of a neurons 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. In this sense, firing is a deviation from that default, normal, inertial state. Several authors[11] have thought that this distinction, between *default, normal*, or *inertial* states and *deviations* therefrom, must be incorporated into a theory of causation. 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*, *non-interial* states 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 $\mathcal{D}$, which provides information about which variable values are default, and which are deviant.[12]

No theory of causation incorporating the default/deviant distinction is complete until it provides an independent characterization of which variable values are default and which are deviant. However, because my focus here will be on simple neuron networks, the only assumption about this distinction I will need is that a neuron's remaining dormant is default, and its firing is deviant.[13]

## 2 Model Invariance

Not all causal models are correct. A causal model which says that the presence of rain is causally determined by the state of my umbrella is not correct; it gets

[11] See in particular Kahneman & Miller (1986), Thomson (2003), McGrath (2005), Maudlin (2004), Hall (2007), Hitchcock (2007), Halpern (2008, 2016), Hitchcock & Knobe (2009), Paul & Hall (2013), and Halpern & Hitchcock (2015).

[12] Formally, we can understand $\mathcal{D}$ as a function from the variables in $\mathcal{U} \cup \mathcal{V}$ to subsets of their values—intuitively, the subset of values which are *default*. The deviant values of $V$ are then all those which are *not* in $\mathcal{D}(V)$. Perhaps this is not enough information—Halpern (2008, 2016) and Halpern & Hitchcock (2015) use a ranking function to represent different *grades* of deviancy. Since my focus here will be on simple neuron systems which have only two states—firing and remaining dormant—I will ignore such complications. In addition to ranking grades of deviancy, we may also have to recognize distinctions in the *direction* or the *valence* of deviancy. For instance, in moral cases, it is natural to think of both the wrong and the supererogatory as departures from the norm, but they are importantly different kinds of departures. Perhaps which variable values are inertial and which are non-inertial 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 a departure from inertial behavior. 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 all of 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] For attempts to characterize the distinction, see Kahneman & Miller (1986), Maudlin (2004), McGrath (2005), Hall (2007), Hitchcock (2007), Hitchcock & Knobe (2009), and Wolff (2016).

the causal structure of the world backwards. Amongst 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 determined 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, or which are causally intermediate between the match's striking and its lighting.

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. 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 it said whether $C$ caused $E$ relative to a single causal model, and if its verdicts did not change from correct model to correct model. That is, it would be nice if our theory satisfied the following constraint.[14]

**Model Invariance** For any two causal models $\mathcal{M}$ and $\mathcal{M}^\dagger$ which both contain the variables $C$ and $E$, if both $\mathcal{M}$ and $\mathcal{M}^\dagger$ are correct, then $C = c$ caused $E = e$ in $\mathcal{M}$ iff $C = c$ caused $E = e$ in $\mathcal{M}^\dagger$.

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 there's oxygen in the room is correct. But, so long as the 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's struck—would be correct, also. Or consider the neuron network displayed in figure 4. The canonical causal model of this neuron network, $\mathcal{M}_4$, includes a variable for $A, C$, and $E$, (with 1 corresponding to firing and 0 corresponding 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 $\mathcal{M}_4$ is correct; but,

[14] There are alternatives to accepting **Model Invariance**. In general, a theory of causation formulated with causal models will specify 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 $C = c$ causing $E = e$ (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 causal models are witnesses to $C = c$ causing $E = e$ (and therefore, $C = c$ didn't cause $E = e$ iff some correct causal model fails to witness $C = c$ causing $E = e$). The first alternative makes it easy to establish causation but difficult to establish non-causation (we must establish non-causation in all of the correct causal models). Likewise, the second makes it easy to establish non-causation, but difficult to establish causation. Model-invariance makes it easy to establish causation and non-causation both.

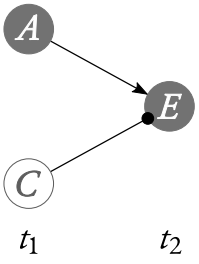



FIGURE 4: *Omission*

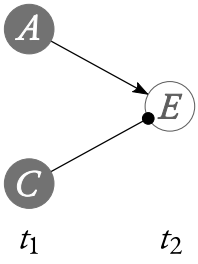



FIGURE 5: *Prevention*

so long as $C$ doesn't fire, the variable for $C$ isn't necessary. Just as we can take the presence of oxygen for granted, so too can we take the non-firing of $C$ for granted. So we can pluck the variable $C$ out of this model and replace it with its actual value, 0, in the structural equation. We will be left with a model—call it '$\mathcal{M}_4^{-C}$'—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 $\mathcal{M} = (\mathcal{U}, \mathbf{u}, V, \mathcal{E}, \mathcal{D})$ is a causal model with the exogenous variable $U \in \mathcal{U}$, then let $\mathcal{M}^{-U}$ be the model that you get by (a) removing $U$ from $\mathcal{U}$; (b) removing $U$'s value from $\mathbf{u}$; (c) 'exogenizing' any variables in $\mathcal{V}$ whose only parent was $U$;[15] (d) replacing $U$ with its value in every structural equation in $\mathcal{E}$; and (e) removing default information about $U$ from $\mathcal{D}$.

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 network in figure 5. This is just like the neuron network from figure 4, except that, in figure 5, $C$ fires, and therefore, $E$ doesn't. The canonical causal model of this neuron network, $\mathcal{M}_5$, will be exactly like $\mathcal{M}_4$, except that the exogenous assignment will tell us that $C = 1$, rather than $C = 0$. This difference 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 $\mathcal{M}_5$, 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 that $E$ and $A$ are causally independent, which is not true. So the model $\mathcal{M}_5^{-C}$ is not correct, even though $\mathcal{M}_5$ is. So removing an exogenous variable does not always preserve correctness.

In general, in order for a structural equation $V := \phi_V(\mathbf{PA}(V))$ to be correct, $\phi_V$ must be a *surjective* function. 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 $\phi_V$. In order to be correct, a structural equation for $V$ must tell us in what circumstances $V$ would take on each of its values. If $\phi_V$ is not surjective, then the structural equation for $V$ can-

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

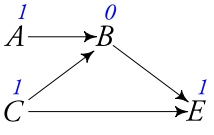
not tell us this. So, if $\phi_V$ is not surjective, then the structural equation for $V$ cannot be correct.

In general, if $U$ is exogenous in $\mathcal{M}$, and if not every structural equation in $\mathcal{M}^{-U}$ is surjective, then say that $U$ is an *essential* exogenous variable in $\mathcal{M}$. And if every structural equation in $\mathcal{M}^{-U}$ is surjective, then say that $U$ is an *inessential* exogenous variable in $\mathcal{M}$. Though removing essential exogenous variables will not preserve correctness, I believe that removing *inessential* exogenous variables always will. That is, I believe we should endorse the following principle.

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

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, *etc*. 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 $\mathcal{M}_1$, 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 $D$ is not necessary. We could pluck the variable $D$ out of this model by replacing it with the left-hand-side of its structural equation, $C$, wherever it appears. We will be left with a model—call it '$\mathcal{M}_1^{-D}$'—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 network, but it tells us about the causal determination structure amongst the variables $A, B, C$, and $E$, and what it tells us about them is all correct.

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

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 structural equations. Those variables are essential. But they are not the only essential endogenous

variables. Consider again the model $\mathcal{M}_1^{-D}$. If we pluck the variable $B$ out of this model in the manner specified above, we will arrive at a model, $\mathcal{M}_1^{-D,-B}$, 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 causal model treats the variables $A$ and $C$ symmetrically. So any theory of causation, presented with the model $\mathcal{M}_1^{-D,-B}$, will say that $C = 1$ caused $E = 1$ iff $A = 1$ caused $E = 1$. But $C = 1$ caused $E = 1$, while $A = 1$ didn't. The causal model $\mathcal{M}_1^{-D,-B}$ is not correct. It tells us that $A$ and $C$ causally determine the value of $E$ along non-intersecting paths, which is not true.

Suppose that, in $\mathcal{M}$, $V$ has a single parent, $Pa$, and at most one child, $Ch$,

$$Pa \to V \to Ch$$

and suppose that $Pa$ is not *also* a parent of $Ch$. If that's so, then say that $V$ is an *interpolated* variable in $\mathcal{M}$. If $V$ is interpolated and all of the equations in $\mathcal{M}^{-V}$ are surjective, then I'll say that $V$ an *inessential* endogenous variable. Through 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 Reduction** If a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$ is correct, and $V \in \mathcal{V}$ is inessential, then $\mathcal{M}^{-V}$ is also correct.

Let's call a theory of causation which is consistent with the principles **Model Invariance**, **Exogenous Reduction**, and **Endogenous Reduction** a *model-invariant* theory of causation. If the theory is inconsistent with the conjunction of these three principles, then let's say that it's 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 causal verdicts changing as we attend to additional variables lying along, or feeding into, the path from $C$ to $E$. Unfortunately, none of the extant theories of causation situated in the framework of causal modelling are model-invariant. In particular, the accounts of Hitchcock (2001, 2007), Halpern & Pearl (2001, 2005), Woodward (2003), Halpern (2008, 2016), and Weslake (forthcoming) are all model-variant.[16]

In the remaining sections, I will develop a theory of causation which is model-invariant. If this theory says that $C = c$ caused $E = e$ in a causal model $\mathcal{M}$, then it will continue to say so after an inessential variable has been removed from $\mathcal{M}$. And, if the theory says that $C = c$ *didn't* cause $E = e$ in a causal model $\mathcal{M}$, then it will continue to say so after an inessential variable has been removed from $\mathcal{M}$. I will introduce this theory by walking through some standard cases from

[16] I show this in a companion paper.

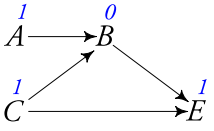
the literature—preemptive overdetermination (§3), counterexamples to transitivity (§4), preemptive prevention (§5), and, finally, symmetric overdetermination (§7).

## 3 Preemptive Overdetermination

Cases of preemptive overdetermination like Tax Cut (§1.2) or the neuron network shown in figure 1 serve as counterexamples to a simple counterfactual theory of causation which says that counterfactual dependence is both necessary and sufficient for causation. Consider the canonical causal model of the neuron network from figure 1, $\mathcal{M}_1$. 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, and $E$ would have fired all the same. (In the counterfactual model $\mathcal{M}_1[C \rightarrow 0]$ in which we intervene so as to set $C$'s value to 0, $E$ 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 of 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 $\mathcal{M}_1$, $C$'s firing caused $E$'s firing. 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 $\mathcal{M}_1$, in the manner described in the previous section. We will then get a causal model, $\mathcal{M}_1^{-D}$, 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 $\mathcal{M}_1$, a Lewisian theory will say that $C = 1$ caused $E = 1$, given the model $\mathcal{M}_1^{-D}$, it will say that $C = 1$ *didn't* cause $E = 1$. So the theory will be model-variant.

Note also that $\mathcal{M}_1^{-D}$ is the canonical causal model of the neuron network shown in figure 6. So if we accept that the canonical causal model is correct, and we want our theory to say that $C$'s firing caused $E$'s firing in figure 6, then we will need another solution to the problem of preemptive overdetermination.

The treatment of preemptive overdetermination favored by almost every author in the causal modeling literature[17] appeals to the variable $B$. Though $E = 1$

[17] See, in particular, Halpern & Pearl (2001, 2005), Hitchcock (2001, 2007), Woodward

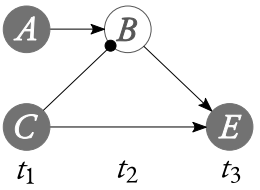
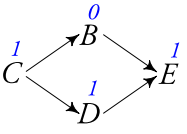



FIGURE 6: *Preemptive Overdetermination*

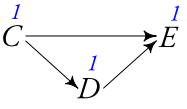
does not counterfactually depend upon $C = 1$ in the model $\mathcal{M}_1$, it *does* counterfactually depend upon $C = 1$ in the *counterfactual* model where we've intervened so as to hold $B$ fixed at its actual value of 0—that is, $\mathcal{M}_1[B \rightarrow 0] \models C = 0 \; \square\!\!\rightarrow E = 0$. And according to these authors, counterfactual dependence in a counterfactual model like this is sufficient to show that $C = 1$ caused $E = 1$. No solution which appeals to the variable $B$ in this way will be model-invariant. For note that the exogenous variable $A$ is inessential in $\mathcal{M}_1$. So, by **Exogenous Reduction**, we may pluck it out, and we will be left with a model, $\mathcal{M}_1^{-A}$, in which the endogenous variable $B$ is (now) inessential.

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




Since $B$ is inessential, we may pluck it out, and we will be left with a model, $\mathcal{M}_1^{-A,-B}$, in which the variable $B$ does not appear at all.

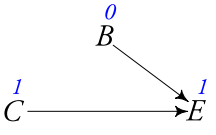
$$E := \neg C \vee D$$
$$D := C$$




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 variable $B$.

Return to the neuron network in figure 6, and consider its canonical causal model, $\mathcal{M}_6$ (which is the same as the model $\mathcal{M}_1^{-D}$, discussed above). 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 this local model, 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 $C = 1$.

(2003), HALPERN (2008, 2016), and WESLAKE (forthcoming). See YABLO (2002, 2004) for similar ideas.

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

$$\mathcal{M}(E) \models C = c^* \ \square\!\!\rightarrow E = e^*$$

In contrast, if there is some $c^* \neq c$ and some $e^* \neq e$ such that

$$\mathcal{M} \models C = c^* \ \square\!\!\rightarrow E = e^*$$

then I will say that $E = e$ *globally* counterfactually depends upon $C = c$ in the model $\mathcal{M}$.[18]

A preliminary proposal, then, is that local dependence suffices for causation. (This is only a *preliminary* proposal—I will want to qualify this claim later on.) This will allow us to say that $C$'s firing caused $E$ to fire in the neuron network shown in figure 6, but it will not allow us to say the same about the neuron network shown in figure 1—for in the canonical model $\mathcal{M}_1$, $E$'s firing neither globally nor locally counterfactually depends upon $C$'s firing (the variable for $C$ is not even included in the local model $\mathcal{M}_1(E)$). I believe that we should handle this case roughly as Lewis (1973) did: by taking causation to be, not dependence, but rather the *transitive closure* of dependence. However, there are a number of counterexamples to the thesis that a chain of dependence is sufficient for causation. Let us turn to those counterexamples now.

## 4 Causal Paths

Sometimes, we can trace out a sequence of events such that each event in the sequence depends upon its predecessor, and thereby conclude that the event at the start of the sequence causes the one at the end. When can we do this? Lewis gave the answer 'always'. This answer allowed him to deal with cases of 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. The answer to give is 'sometimes, but not always', and

[18] Throughout, I am presupposing that $\mathcal{M} \models C = c \wedge E = e$.


FIGURE 7: The octogonal neurons can either fire weakly (light grey) or strongly (dark grey). The square-headed connection between $C$ and $B$ tells us that, if $C$ fires, this diminishes the strength with which $B$ would otherwise have fired. In figure 7a, $E$'s firing depends upon $B$'s firing weakly. And $B$'s firing weakly depends upon $C$'s firing. But $C$'s firing didn't cause $E$ to fire.

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 $E$,

$$C \rightarrow D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_N \rightarrow E$$

is what I will call a *causal path*.[19] If there is a causal path running from $C$ to $E$ in a causal model, then we may conclude that $C$'s value caused $E$'s value. The LEWISIAN view is that a directed path is causal whenever the value of each variable in the path depends upon the value of its parent on the path. I believe that we should impose additional constraints on a path being causal. I'll introduce these constraints by surveying some representative counterexamples to the LEWISIAN view.

### 4.1 CAUSAL PATHS AND CONTRASTS

One class of counterexamples to the LEWISIAN view is well illustrated by the neuron network illustrated in figure 7.[20] In this neuron network, the octogonal neurons $A$ and $B$ are special. They can either fire *weakly* (indicated with light grey coloring) or *strongly* (indicated with dark grey). The square-headed 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, *e.g.*, 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, no

[19] A *directed path* is a sequence of directed edges $V_i \rightarrow V_{i+1}$ such that $V_i \in \mathbf{PA}(V_{i+1})$, for each $i$. Throughout, whenever I say 'path', I should be understood to be talking about a *directed* path.

[20] *Cf.* PAUL & HALL (2013), and also LEWIS (1986, p. 210).

matter 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 network provides a counterexample to the Lewisian view that causation is the transitive closure of dependence.

For another case with the same causal 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.[21]

I follow Maslen (2004) and Schaffer (2005) in thinking that cases like these illustrate the importance of paying attention to *contrasts* in chains of dependence.[22] 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 $C = c^* \,\square\!\!\rightarrow\, E = e^*$; the second, (b), is made true by a counterfactual $C = c^* \,\square\!\!\rightarrow\, E = e^{**}$; and the third, (c), is made true by a counterfactual $C = c^{**} \,\square\!\!\rightarrow\, E = e^*$.[23] The lesson of figure 7 is this: in order for a path to be causal, it is not enough that the value of each variable along the path depend upon the value of its parent. The relevant contrasts along the path also have to match up.

A bit of terminology: given a variable, $V$, on a directed path, $\mathcal{P}$, let's call the parent of $V$ on $\mathcal{P}$ $V$'s $\mathcal{P}$-parent.[24] Then, as a preliminary account, we may say:

> Causal Path (provisional)
> A directed path $\mathcal{P} : C \rightarrow D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_N \rightarrow E$ is a *causal path* leading from $C$ to $E$ only if there is an assignment of contrasts to the variables along $\mathcal{P}$ such that:
>
> (a) starting with $D_1$, each variable value along the path, rather than its contrast, depends upon its $\mathcal{P}$-parent's value, rather than its contrast.

And our preliminary theory is that, if there is a causal path leading from $C$ to $E$,

---

[21] See McDermott (1995), as well as the counterexamples to transitivity discussed in Paul (2004).

[22] See also Hitchcock (1996) and Schaffer (2012a) for more on contrasts in causal claims.

[23] Of course, in order for these dependence claims to be true, it must also be that $C = c \wedge E = e$. Throughout, I am using '$c$' and '$e$' for the actual values of $C$ and $E$. Also, though I am not bothering to say it explicitly, it should be understood throughout that $c^* \neq c$ and $e^* \neq e$.

[24] Note: in $\mathcal{M}_6$, relative to the path $\mathcal{P} : C \rightarrow B \rightarrow E$, $C$ is a parent of $E$ on the path $\mathcal{P}$, but it is not its $\mathcal{P}$-parent. Its $\mathcal{P}$-parent is $B$.

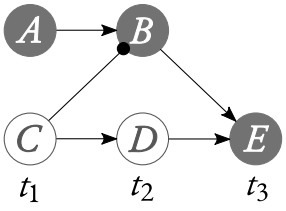



FIGURE 8

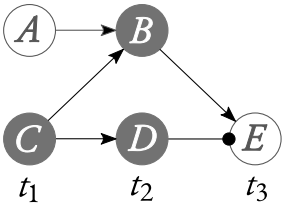



FIGURE 3

then $C = c$ caused $E = e$, even if there's no direct dependence between $E = e$ and $C = c$.

### 4.2 Causal Paths, 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.[25] Consider again the neuron network of *preemptive overdetermination* from figure 1, but suppose that $C$ doesn't fire, as in figure 8. In this neuron network, $E$'s firing depends upon $B$'s firing (rather than not). And $B$'s firing (rather than not) depends upon $C$'s failure to fire. So we have a chain of dependence with matching contrasts leading from $C$ to $E$, but $C$'s failure to fire didn't cause $E$ to fire.

Or consider again the neuron network from figure 3 (reproduced above). 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 firing did not cause $E$ to remain dormant.

As we've already seen (§1.2), were it not for the information about which variable values are default and which are deviant, we could model the neuron network in figure 3 with a model isomorphic to the canonical causal 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 default information. Note also that **Exogenous Reduction** and **Endogenous Reduction** allow us to remove every variable other than $C$ and $E$ from $\mathcal{M}_3$. $A$ is inessential, so **Exogenous Reduction** tells us that the model $\mathcal{M}_3^{-A}$ is correct. In the model $\mathcal{M}_3^{-A}$, $B$ is inessential, so **Endogenous Reduction** tells us that the model $\mathcal{M}_3^{-A,-B}$ is correct. And similarly, in the model $\mathcal{M}_3^{-A}$, $D$ is inessential, so **Endogenous Reduction** tells us that the model $\mathcal{M}_3^{-A,-D}$ 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

[25] SCHAFFER is working in a different theoretical framework; and it affords him a response to the kinds of counterexamples raised below (see p. 342).

theory should not depend upon the default information of any variables other than $C$ and $E$ themselves.[26]

In both figure 8 and figure 3, it is noteworthy that either $C$ or $E$ takes on a default, normal or inertial value (rather than a default value). Whereas, in figure 1, both $C$ and $E$ take on deviant, abnormal, non-inertial values. This suggests that, in order for a path to be causal, the values of the variables at the start and end of the path must be deviant, rather than default. Let's add this to our account. A path is causal when (a) the value of each variable along the path, rather than its contrast, depends upon the value of its parent on the path, rather than its contrast; and (b) the variable values at the start and end of the path are deviant, with default contrasts.

> Causal Path (provisional)
> A directed path $\mathcal{P} : C \to D_1 \to D_2 \to \cdots \to D_N \to E$ is a *causal path* leading from $C$ to $E$ only if there is an assignment of contrasts to the variables along $\mathcal{P}$ such that:
>
> (a) starting with $D_1$, each variable value along the path, rather than its contrast, depends upon its $\mathcal{P}$-parent's value, rather than its contrast; and
>
> (b) $C$ and $E$ have deviant values, with default contrasts.

If we suppose that survival is an inertial state—the state in which people normally remain unless they are acted upon from without—then clause (b) explains why the boulder's becoming dislodged does not cause Matthew to survive (§1.2), 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 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.[27]

### 4.3 Causal Paths and Switching

Consider the neuron network shown in figure 9. There, the neurons $A, B, C, D$, and $F$ are just as in the short circuit from figure 3. In the canonical causal model

[26] 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 $\mathcal{M}_3^{-A,-B}$, and $B$ is not inessential in $\mathcal{M}_3^{-A,-D}$. So for all we've said, it could be that the default information about $D$ should be relevant to the theory's verdicts in $\mathcal{M}_3^{-A,-B}$, while the default information about $B$ should be relevant to the theory's verdicts in $\mathcal{M}_3^{-A,-D}$. So while these considerations give us strong reason to suspect that a model-invariant account will have to appeal only to the default information of $C$ and $E$ themselves, they are not apodeictic.

[27] If we suppose that wrong acts are categorized as deviant, while right acts are categorized as default, then it also explains cases like *Shock C* from McDermott (1995). (Though I suspect that properly treating cases like these in general will require more thought about the distinction between inertial and non-inertial states—see footnote 12.)

$\mathcal{M}_9$, $F$'s value is default, so the path $C \to D \to F$ is not causal, and our provisional account will not tell us that $C$'s firing caused $F$ to not fire. However, since $E$'s value *is* deviant, the path $C \to D \to F \to E$ *will* be causal. $E$'s firing (rather than not) depends upon $F$'s remaining dormant (rather than firing). $F$'s remaining dormant (rather than firing) depends upon $D$'s firing (rather than not). And $D$'s firing (rather than not) depends upon $C$'s firing (rather than not). So there's a path of dependence leading from $C$ to $E$, the contrasts along the path match, and the variable values at the beginning and end of the path are deviant, rather than default. So the provisional account tells us that $C$'s firing caused $E$ to fire in the neuron network in figure 9. This is a problem, since $C$'s firing did not cause $E$ to fire in figure 9. The only reason we might think $C$'s firing caused $E$ to fire is that it did so by causing $F$ to remain dormant. But $C$'s firing didn't cause $F$ to remain dormant.

Or consider the neuron network shown in figure 10. There, the neurons $F, B, D$ and $E$ are just like $C, B, D$, and $E$ in figure 8. In the canonical causal model $\mathcal{M}_{10}$, $F$'s value is default, so the path $F \to B \to E$ is not causal, and our provisional account will not tell us that $F$'s failure to fire caused $E$ to fire. However, since $C$'s value *is* deviant the path $C \to F \to B \to E$ *will* be causal. $E$'s firing (rather than not) depends upon $B$'s firing (rather than not). $B$'s firing (rather than not) depends upon $F$'s remaining dormant (rather than firing). And $F$'s remaining dormant (rather than firing) depends upon $C$'s firing (rather than not). So there's a path of dependence leading from $C$ to $E$, the contrasts along the path match, and the variable values at the beginning and end of the path are deviant, rather than default. So the provisional account tells us that $C$'s firing caused $E$ to fire in figure 10. This is a problem, since $C$'s firing did not cause $E$ to fire. In figure 10, $C$ is a 'switch'. If $A$ fires, it will precipitate a chain of neuron firings leading to $E$, whether $C$ fires or not. If $C$ fires, then $A$'s signal will take the top path, through $B$ to $E$. If $C$ doesn't fire, then $A$'s signal will take the bottom path, through $F$ and $D$ to $E$. While $C$'s firing caused $A$'s signal to take the top path, $C$'s firing did not cause $E$ to fire.[28]

Suppose we have a directed path $\mathcal{P} : C \to D_1 \to D_2 \to \cdots \to D_N \to E$, and along this path lie two variables, $D$ and $R$. If our model has *another* directed path from $D$ to $R$, $\mathcal{O} : D \to O_1 \to O_2 \cdots \to O_M \to R$ (think of the $O_i$ as variables *off* the path $\mathcal{P}$), such that none of the directed edges from $\mathcal{O}$ appear in $\mathcal{P}$, then say that $D$ is a *departure* variable, and say that $R$ is one of its *return* variables (relative to the path $\mathcal{P}$). For instance, in the model $\mathcal{M}_9$, relative to the path $C \to D \to F$, $C$ is a departure variable, and $F$ is its return. Likewise, in the model $\mathcal{M}_{10}$, relative to the path $C \to F \to B \to E$, $F$ is a departure variable, and $E$ is its return; and, relative to the path $A \to B \to E$, $A$ is a departure variable with returns $B$ and $E$. In contrast, relative to the path $D \to F \to E$ in $\mathcal{M}_9$, $F$ is not a return variable.

[28] See Hall (2004), Sartorio (2005), and Paul & Hall (2013, p. 232–233).

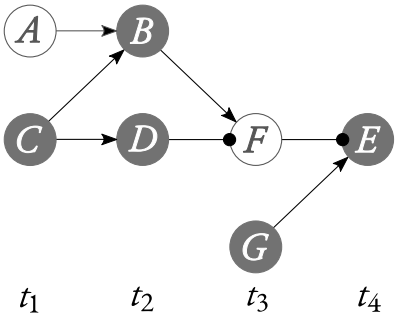



Figure 9

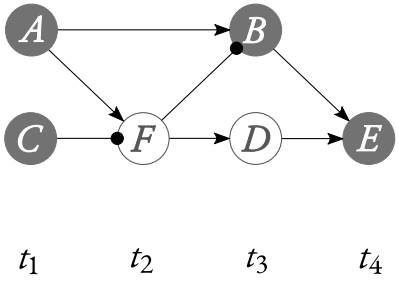



Figure 10

And, relative to the path $C \rightarrow F \rightarrow B$ in $\mathcal{M}_{10}$, $F$ is not a departure variable.

Take some path, $\mathcal{P}$, with a departure-return variable pair, $< D, R >$. If $\mathcal{P}$ is to be causal, then $D$ must be assigned some contrast value—call it '$d^*$'. Let $\mathbf{I}$ be the intermediate variables between $D$ and $R$ on $\mathcal{P}$ (excluding $D$ and $R$ themselves), and let $\mathbf{i}^*$ be their designated contrasts. Then, we will say that $< D, R >$ is an *active* departure-return variable pair (relative to the path $\mathcal{P}$ and an assignment of contrasts to the variables on $\mathcal{P}$) if some parent of $R$ which is not on $\mathcal{P}$ takes on a different value in the model $\mathcal{M}[\mathbf{I} \rightarrow \mathbf{i}^*, D \rightarrow d^*]$ than it does in the model $\mathcal{M}[\mathbf{I} \rightarrow \mathbf{i}^*]$.[29] If $< D, R >$ is an active departure-return variable pair, then, when the variables between $D$ and $R$ on the path are held fixed at their contrast values, wiggling $D$ wiggles some non-$\mathcal{P}$ parent of $R$. In that case, $D$ potentially affects $R$ both along $\mathcal{P}$ and along some other path or paths. It could be that, what $D$ gives $R$ along one path, it takes away along the others. If $D$ gives a deviant value to $R$ along $\mathcal{P}$—that is, if $D$ and $R$ both take on deviant, rather than default, values (as with $C$ and $E$ along the path $C \rightarrow D \rightarrow E$ in the case of preemptive overdetermination from figure 1)—then this will make no difference with respect to whether $\mathcal{P}$ is a causal path. But if $D$ and $R$ are active, and $D$ does not give a deviant value to $R$, then $\mathcal{P}$ is not causal.[30] In order for a directed path to count as a causal path, it must be that, for every active departure-return variable pair $< D, R >$, both $D$ and $R$ have deviant values and default contrasts.

In summary, a directed path is causal iff (a) each variable value, rather than its contrast, depends upon its parent on the path, rather than its contrast; (b) the first and last variables take on deviant, rather than default, values; and (c) every active departure, return variable pair takes on deviant, rather than default, values.

---

[29] If $D$ is $R$'s $\mathcal{P}$-parent, then $D$ and $R$ are active iff some non-$\mathcal{P}$ parent of $R$'s takes on a different value in $\mathcal{M}[D \rightarrow d^*]$ than it does in $\mathcal{M}$.

[30] Couldn't an active departure $D$ have a default value or a deviant contrast, yet still not take away along other paths what it gives $R$ along $\mathcal{P}$? Yes, though we'll have to wait until §7 and the introduction of *forking* causal paths to handle these kinds of cases.

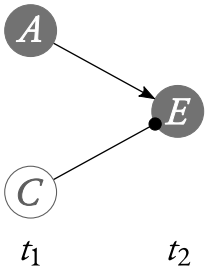



FIGURE 4: *Omission*

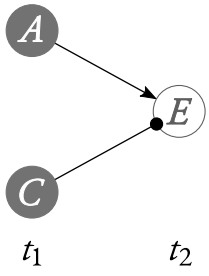



FIGURE 5: *Prevention*

CAUSAL PATH (PROVISIONAL)
A directed path $\mathcal{P} : C \rightarrow D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_N \rightarrow E$ is a *causal path* leading from $C$ to $E$ if and only if there is an assignment of contrasts to the variables along $\mathcal{P}$ such that:

(a) starting with $D_1$, each variable value along the path, rather than its contrast, depends upon its $\mathcal{P}$-parent's value, rather than its contrast.

(b) $C$ and $E$ have deviant values and default contrasts.

(c) for every active departure-return variable pair along $\mathcal{P}$, $< D, R >$, $D$ and $R$ have deviant values and default contrasts.

The account is still provisional—not because additional conditions need to be added, but rather because more must be said about the kind of dependence mentioned in clause (a) of CAUSAL PATH. Is it local dependence? Global dependence? Will either suffice? The answer to this question won't make a difference to any of the cases from this section; however, it will make a difference to our verdicts about the case of preemptive overdetermination from figure 6, as well as putative cases of prevention and omission without dependence (§5). We will return to the question of which kind of dependence to appeal to in clause (a) of CAUSAL PATH at the end of §5.

## 5 PREVENTION AND OMISSION WITHOUT DEPENDENCE?

Thus far, all the examples of causation we have considered have been cases where both the cause and effect variables take on deviant values. However, many theorists believe that default variable values can be both causes and effects. For instance, consider the case of omission from figure 4 (reproduced here). If global dependence suffices for causation, then $C$'s failure to fire caused $E$ to fire. For, had $C$ fired, $E$ would not have. Since several theorists say that global dependence suffices for causation, they say that default variable values can be causes. Or consider the case of prevention from figure 5. If global dependence suffices for causation, then $C$'s firing caused $E$ to not fire. For, had $C$ not fired, $E$ would

have. So these theorists say that default variable values can be effects. While some are happy with these verdicts, others see cases of prevention and omission as reasons to doubt that global dependence is sufficient for causation.[31] I will assume for the moment that global dependence suffices for causation—but I'll return to this question in §6.

We've granted that global dependence suffices for causation, even when the cause or effect variable takes on a default value. Does *local* dependence suffice for causation, when the putative effect is default? This question turns out to be closely related to the question of whether there are cases of *preemptive prevention*, or preemption without global dependence.[32] For instance, consider the neuron network in figure 11a. Either $A$ or $C$'s firing would be sufficient, on its own, to prevent $E$ from firing; but, as it turns out, both fired. Did $C$'s firing cause $E$ to not fire? There is an initial temptation to say 'no'. This initial temptation can fade if you are asked, 'of $A$ and $C$, which one prevented $E$ from firing?'. It is then tempting to reason as follows: 'well, $A$ and $C$ are the only candidate causes of $E$'s not firing, and clearly $A$ didn't cause $E$ to not fire; so it must have been $C$ that did it'.[33]

We could secure the verdict that $C$'s firing caused $E$ to not fire in figure 11a if we allowed local dependence to suffice for causation, even when the effect variable takes on a default value. For take the canonical causal model of this neuron network, $\mathcal{M}_{11a}$. The local model at $E$, $\mathcal{M}_{11a}(E)$, will contain the sole structural equation $E := F \wedge \neg(C \vee B)$, and the exogenous assignments $B = 0$ and $C = F = 1$. This local model entails that, were $C$ to take on the value 0, $E$ would take on the value 1. So there is local dependence between $E$'s remaining dormant and $C$'s firing.

It's not immediately clear whether we should want our theory to deliver this verdict or not. What is clear is that, whatever our theory says about whether $C$'s firing caused $E$ to remain dormant in figure 11a, it should say the very same thing about the neuron network in figure 11b. However, in the canonical causal model of figure 11b, $\mathcal{M}_{11b}$, there is no local dependence between $E$ and $C$—$C$ isn't even in the local modal at $E$, $\mathcal{M}_{11b}(E)$. So, if we wish to say that $C$'s firing caused $E$ to remain dormant in figure 11b, we would have to appeal to a causal path running from $C$ to $E$. After our conclusions in the previous section, this option is not available. $C \to D \to E$ is not a causal path, since $E$'s value is not deviant. So, in $\mathcal{M}_{11b}$, we cannot say that $C = 1$ caused $E = 0$. And if we cannot say this

---

[31] See, *e.g.*, Beebee (2004) and McGrath (2005). Others treat cases of omission and prevention as second-class causal citizens—their status as causal is derivative or parasitic upon the more paradigm case of production. See *e.g.*, Dowe (2000) and Hall (2004).

[32] See McDermott (1995) and Collins (2004).

[33] This reasoning is compelling, but fallacious. We can accept that $A$ and $C$ *jointly* caused $E$ to not fire without thinking that either of them did it individually—see §7.

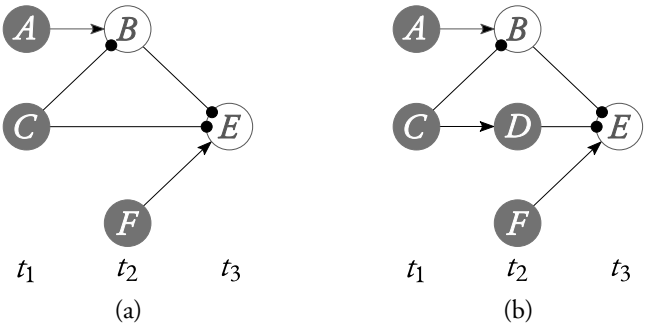
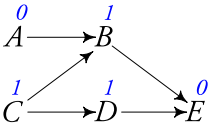



FIGURE 11: *Prevention without Dependence?*

about the neuron network in figure 11b, then we shouldn't say it about the neuron network in figure 11a, either. So we should not allow local dependence to suffice for causation when the effect variable is default.

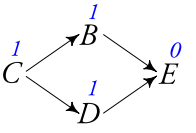
We can further support this conclusion by noting that any model-invariant theory of causation will say that $C$'s firing caused $E$ to not fire in figure 11b iff it says that $C$'s firing caused $E$ to not fire in the 'short circuit' from figure 3. Begin with the canonical causal model of the neuron network from figure 3, $\mathcal{M}_3$,

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



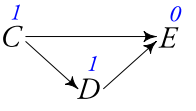

In this model, the exogenous variable $A$ is inessential. So we may pluck it out, leaving behind the model $\mathcal{M}_3^{-A}$,

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



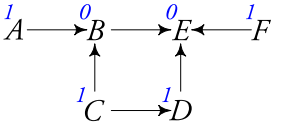

And, in this model, the interpolated endogenous variable $B$ is inessential. So we may remove it, leaving behind the model $\mathcal{M}_3^{-A,-B}$,

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




And we may arrive at an isomorphic causal model by beginning with the canonical causal model of the putative case of preemptive prevention from figure 11b, $\mathcal{M}_{11b}$, and pruning inessential variables. In the canonical model,

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



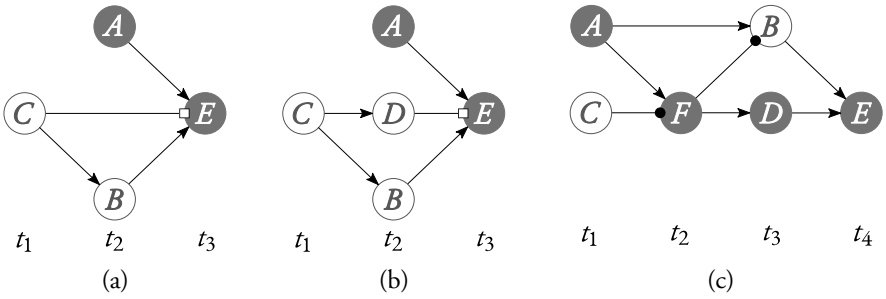
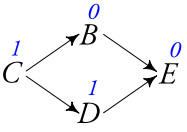



FIGURE 12: *Omission without Dependence*? If the neuron at the base of a square-headed connection fires, this will cancel out any one stimulatory signal coming into the neuron at its head. Thus, in figure 12a, $E$ will fire iff either $C$ doesn't and either $A$ or $B$ does, or $C$ does, and both $A$ and $B$ do.

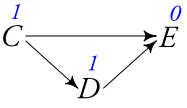
the exogenous variables $A$ and $F$ are inessential, so **Exogenous Reduction** tells us that, if $\mathcal{M}_{11b}$ is correct, then so too is $\mathcal{M}_{11b}^{-A,-F}$,

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




And, in this model, the endogenous variable $B$ is inessential, so **Endogenous Reduction** tells us that, if $\mathcal{M}_{11b}^{-A,-F}$ is correct, then so too is $\mathcal{M}_{11b}^{-A,-F,-B}$,

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




But $\mathcal{M}_{11b}^{-A,-F,-B}$ is isomorphic to the model $\mathcal{M}_{3}^{-A,-B}$ of the 'short circuit' from figure 3. So a model-invariant theory of causation will say that $C$'s firing caused $E$ to not fire in figure 11b iff it says the same about figure 3. Since we should not say that $C$'s firing caused $E$ to not fire in figure 3, we should not say that figure 11b provides a case of prevention without dependence.[34,35]

Similar reasoning points us towards the conclusion that local dependence doesn't suffice for causation when the putative *cause* variable is default. Consider the putative case of omission without dependence shown in figure 12a. There may be some temptation to say that $C$'s failure to fire caused $E$ to fire, even though,

[34] Does this mean that nothing caused $E$ to remain dormant? Not necessarily. For we may say that $A$ and $C$ *jointly* caused $E$ to remain dormant, without saying that either of them caused it individually. See §7.

[35] This all assumes, of course, that $E$'s failure to fire in figure 11b is *inertial*. But perhaps this assumption is mistaken. Perhaps, given that $F$ fired, $E$'s firing is its inertial state, and its failure to fire is a departure from that inertial behavior (see fn 12). If so, then we should say that $C$'s firing caused $E$ to not fire.

had $C$ fired, $E$ still would have fired (though I expect this temptation to be weaker than in the putative case of preemptive prevention). Once again, we could secure this judgment if we allowed local dependence to suffice for causation, even when the cause variable takes on a default value. Perhaps we should want our theory to deliver this verdict, but it's certainly not clear that we do. What is clear is that we should want our theory to say the same thing about the neuron networks in figures 12a and 12b. However, in the canonical causal model of figure 12b, $\mathcal{M}_{12b}$, there is no local dependence between $E = 1$ and $C = 0$. The variable $C$ doesn't even appear in the local model $\mathcal{M}_{12b}(E)$. So if we wish to say that $C$'s failure to fire caused $E$ to fire in figure 12b, we would have to appeal to a causal path running from $C$ to $E$. However, $C \rightarrow D \rightarrow E$ is not a causal path, since $C$'s value is not deviant. So, in $\mathcal{M}_{12b}$, we cannot say that $C = 0$ caused $E = 1$. And if we cannot say this about the neuron network in figure 12b, then we shouldn't say it about the neuron network in figure 12a, either. So we should not allow local dependence to suffice for causation when the cause variable is default.

Paul & Hall (2013, pp. 187 & 214) suggest that neuron networks like the one shown in figure 12c provide cases of causation by omission without dependence.[36] Had $C$ fired, it would have interrupted the causal process running from $A$ to $E$. Paul & Hall suggest that $C$'s failure to interrupt this process should be counted among the causes of $E$'s firing, even though, had $C$ fired, $E$ would have fired all the same, since $A$ would have caused $E$ to fire through a different process. But figure 12c is just the case of *switching* from figure 10. (Figure 12c shows what would have happened in figure 10, had the switch neuron $C$ not fired.) In figure 12c, $C$'s failure to fire caused the signal from $A$ to take the lower path to $E$, rather than the upper path. But it did not cause $E$ to fire.

With these points in mind, return to the question with which we ended the previous section: what kind of dependence should we allow to compose a causal path? Cases of preemptive overdetermination like the one from figure 6 show us that we should allow local dependence between variables with deviant values and default contrasts. However, putative cases of preemptive prevention/omission have taught us that we should not allow local dependence between variables with default values or deviant contrasts. Let's define a general notion of what we can call *causal dependence*. If both $C$ and $E$ takes on deviant, rather than default, values, then causal dependence is local dependence. If, however, either $C$ or $E$ take on a default value or has a deviant contrast, then causal dependence is global dependence.

> Causal Dependence
> For $C \in \mathbf{PA}(E)$, $E = e$, rather than $e^*$, *causally depends* upon $C = c$,

[36] Paul & Hall use a slightly different neuron network, but any model-invariant account of causation will have to render the same verdict about their neuron network and the one in figure 12c.

rather than $c^*$, iff either (Dev) or (Def).

(Dev) $c$ and $e$ are both deviant, with $c^*$ and $e^*$ default, and $E = e$, rather than $e^*$ *locally* depends upon $C = c$, rather than $c^*$,

$$\mathcal{M}(E) \models C = c^* \; \square\!\!\rightarrow E = e^*$$

(Def) Either $c$ or $e$ is default, or $c^*$ or $e^*$ is deviant, and $E = e$, rather than $e^*$, *globally* depends upon $C = c$, rather than $c^*$,

$$\mathcal{M} \models C = c^* \; \square\!\!\rightarrow E = e^*$$

(Note that causal dependence is only defined between a variable and one of its causal parents. Since the relation of causal parenthood is model-relative, our relations of causal dependence are model-relative, too.)

A causal path should be built from chains of causal dependence. This, then, is our final account of when a directed path is causal:

Causal Path
A directed path $\mathcal{P} : C \rightarrow D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_N \rightarrow E$ is a *causal path* leading from $C$ to $E$ iff there is an assignment of contrasts to the variables along $\mathcal{P}$ such that:

(a) starting with $D_1$, each variable value along the path, rather than its contrast, causally depends upon its $\mathcal{P}$-parent's value, rather than its contrast.

(b) $C$ and $E$ have deviant values, with default contrasts.

(c) for every active departure-return variable pair along $\mathcal{P}$, $< D, R >$, $D$ and $R$ have deviant values, with default contrasts.

## 6 Production, Causation, and Dependence

If there is a causal path leading from $C$ to $E$, then $C = c$ caused $E = e$. In this case, let's call $C = c$ a *productive cause* of $E = e$.

Productive Causation
In a causal model $\mathcal{M}$, $C = c$ is a *productive cause* of $E = e$ if and only if there is a causal path leading from $C$ to $E$.

Productive causation is a model-invariant relation. Suppose we have a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with $U \in \mathcal{U}$. And suppose that $U \neq C, E$ and that $U$ is inessential. Then, $C = c$ will be a productive cause of $E = e$ in $\mathcal{M}$ if and only if $C = c$ is also a productive cause of $E = e$ in $\mathcal{M}^{-U}$. Similarly, if we have a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$ with an inessential $V \in \mathcal{V}$, and $V \neq C, E$, then $C = c$ will be a productive cause of $E = e$ in $\mathcal{M}$ if and only if

$C = c$ is also a productive cause of $E = e$ in $\mathcal{M}^{-V}$. (See **Proposition 1** in the appendix.)

Contrast productive causation with another relation we can call *production*. $C = c$ produces $E = e$ iff there is a *productive path* leading from $C$ to $E$.

> Production
> In a causal model $\mathcal{M}$, $C = c$ *produces* $E = e$ if and only if there is a productive path leading from $C$ to $E$.

A *productive* path is a directed path such that every variable on the path takes on a deviant value, and its taking on that deviant value, rather than some default value, locally depends upon its parent on the path taking on a deviant, rather than a default, value.

> Productive Path
> A directed path $\mathcal{P} : C \rightarrow D_1 \rightarrow D_2 \rightarrow \cdots \rightarrow D_N \rightarrow E$ is a *productive path* leading from $C$ to $E$ iff there is an assignment of contrasts to the variables along $\mathcal{P}$ such that
>
> (a) starting with $D_1$, the value of each variable along the path, rather than its contrast, locally depends upon its $\mathcal{P}$-parents value, rather than its contrast.
>
> (b) every variable along the path has a deviant value, with a default contrast.

In any causal model, $\mathcal{M}$, if $C = c$ produces $E = e$ in $\mathcal{M}$, then $C = c$ will be a productive cause of $E = e$ in $\mathcal{M}$. So producing an effect is one way of productively causing it. But the converse does not hold. $C = c$ can productively cause $E = e$ without producing it.

A productive path is so-called because it provides a natural characterization of the notion of a productive causal process in the terms of causal models.[37] So understood, a productive causal process is an uninterrupted process which locally propagates deviant or non-inertial states. What it is for this deviancy to be *propagated* is for each stage in the process to locally depend upon its predecessor. Notice that $C$'s firing produces $E$'s firing in the canonical models of *preemptive overdetermination* in figures 1 and 6. Similarly, $A$'s firing produces $E$'s firing in the canonical models of figures 4, 8, 10, 12c, and 13. In general, it seems that, if $C$ produces $E$, the judgment that $C$ caused $E$ is intuitive and uncontroversial. There is little debate about whether $C$'s firing caused $E$ to fire in figures 1 and 6, or whether $A$'s firing caused $E$ to fire in figures 4, 8, 10, 12c, and 13.

[37] The notion I am characterizing here is not the notion of a causal process provided by authors like Fair (1979), Salmon (1984, 1994), and Dowe (2000)—those notions are characterized in physical terms, rather than the terms of a causal model—but there are some similarities. See also Hall (2004)'s characterization of causal production.

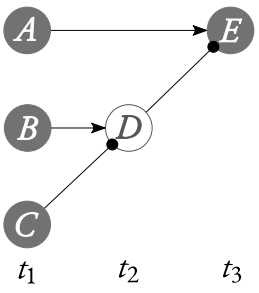



FIGURE 13: *Double Prevention*

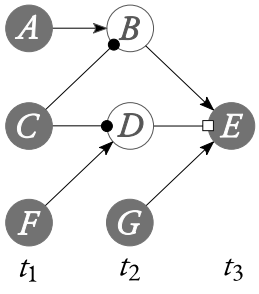



FIGURE 14: *Double Prevention without Dependence*

In contrast, in cases of double prevention like the one shown in figure 13, $C$'s firing does not produce, but does productively cause, $E$'s firing. There, $D$ is a potential preventer of $E$'s firing. $C$'s firing prevents $D$ from preventing $E$ from firing. In the canonical causal model $\mathcal{M}_{13}$, $C \rightarrow D \rightarrow E$ is a causal path, so $C$'s firing is a productive cause of $E$'s firing. However, $C \rightarrow D \rightarrow E$ is not a productive path, since the intermediate variable $D$ takes on a default, rather than a deviant, value.

Or consider the case of *double prevention without dependence* from figure 14. (Recall, if the neuron at the base of a square-headed connection fires, this will cancel out one of the stimulatory signals coming into the neuron at its head. Thus, $E$ will fire iff either $D$ doesn't fire and either $B$ or $G$ does, or $D$ does fire, and both $B$ and $G$ do.) If we ignore the neurons $A$ and $B$, then figure 14 displays another case of double prevention. $D$ is a potential preventer of $E$'s firing, but $C$'s firing prevents $D$ from preventing $E$ from firing. Unlike in figure 13, however, $E$'s firing does not globally depend upon $C$'s firing. For, had $C$ not fired, $B$ would have fired, and $E$ would have fired all the same. $C \rightarrow D \rightarrow E$ is a causal, but not a productive, path leading from $C$ to $E$. So $C$'s firing productively caused, but did not produce, $E$'s firing.

Though I say that $C$'s firing caused $E$ to fire in figures 13 and 14, people's judgments about causation in these cases tend to be less uniform. This suggests that production lies at the heart of our concept of causation. If $C$ produces $E$, then we should expect little disagreement about whether $C$ caused $E$. If, on the other hand, $C$ causes $E$ without producing it, then we should expect more controversy.

If PRODUCTION captures a more intuitive, natural, and uncontroversial notion of causation—if opinion about productive causes which do not produce their effects is less uniform—then why not treat causation as production? Because, unlike productive causation, production is a model-variant relation. Take the canonical causal model of the case of double prevention from figure 13, $\mathcal{M}_{13}$,

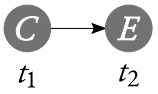
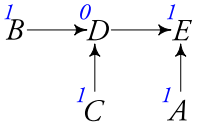


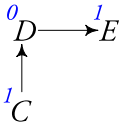

Figure 15

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



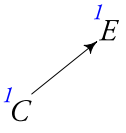

In this model, the exogenous variables $A$ and $B$ are both inessential. So **Exogenous Reduction** tells us that we may remove them both, leaving behind the model $\mathcal{M}_{13}^{-A,-B}$,

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




In this model, the endogenous variable $D$ is inessential, so **Endogenous Reduction** tells us that we may remove it, leaving behind the model $\mathcal{M}_{13}^{-A,-B,-D}$,

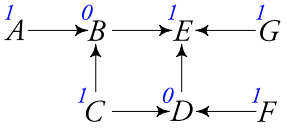
$$E := C$$




But, in this model, $C = 1$ produces $E = 1$—indeed, this is the canonical causal model of the neuron diagram from figure 15, a paradigmatic instance of causal production. So if we understood causation as production, our causal verdicts would change as we attended to additional variables lying along, or feeding into, the path from cause to effect.[38]

Or consider the case of double prevention without dependence from figure 14. We may begin with the canonical causal model $\mathcal{M}_{14}$,[39]

$$E := (B + G) > D$$
$$D := F \wedge \neg C$$
$$B := A \wedge \neg C$$




In this model, the exogenous variables $F$ and $G$ are inessential. So **Exogenous Reduction** tells us that they may be removed, leaving behind the model $\mathcal{M}_{14}^{-F,-G}$,

[38] Schaffer (2000) argues that, in many paradigm instances of causal production—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.

[39] '$E := (B+G) > D$' says that $E$'s value will be determined to be the truth-value of the proposition $(B + G) > D$.

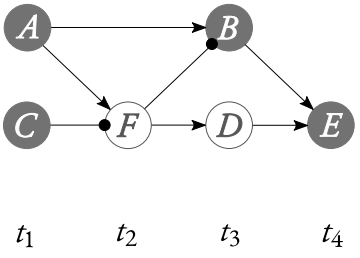
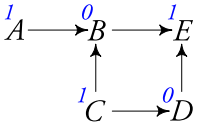


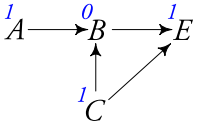

Figure 10

$$E \coloneqq B \vee \neg D$$
$$D \coloneqq \neg C$$
$$B \coloneqq A \wedge \neg C$$




And, in this model, the interpolated endogenous variable $D$ is inessential. So **Endogenous Reduction** tells us that it can be removed, leaving behind the model $\mathcal{M}_{14}^{-F,-G,-D}$,

$$E \coloneqq B \vee C$$
$$B \coloneqq A \wedge \neg C$$

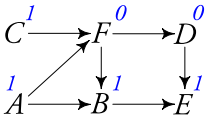



which is the canonical causal model of the case of *preemptive overdetermination* from figure 6. So a model-invariant theory of causation must say that $C$'s firing caused $E$ to fire in figure 14 iff it says that $C$'s firing caused $E$ to fire in figure 6. Since we should want our theory to say the latter, and since we should want our theory to be model-invariant, we should want it to say the former as well.

While productive paths lie at the heart of our concept of causation, the additional latitude in clauses (a) and (b) of Causal Path is needed to secure model-invariance. It thereby helps to guarantee that our causal verdicts do not depend upon whether we have attended to every interpolated lying variable along, or leading into, the path from cause to effect.

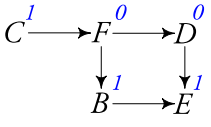
Model-invariance additionally explains the curious clause (c) of Causal Path—the clause which requires active departure and return variables to take on deviant, rather than default, values. Recall from §4.3 that this clause allowed us to capture the intuitive judgment that the 'switch' neuron $C$ in figure 10 (reproduced here) didn't cause $E$ to fire. But this clause does more than just capture the data. Were it not for this clause, productive causation would be a model-variant relation. Take the canonical causal model $\mathcal{M}_{10}$.

$$E \coloneqq B \vee D$$
$$B \coloneqq A \wedge \neg F$$
$$D \coloneqq F$$
$$F \coloneqq A \wedge \neg C$$



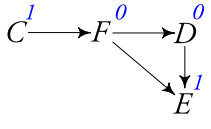

Without clause (c) in Causal Path, $C \rightarrow F \rightarrow B \rightarrow E$ would count as a causal path. For $E$'s firing causally depends upon $B$'s firing, $B$'s firing causally depends upon $F$'s failure to fire, and $F$'s failure to fire causally depends upon $C$'s firing. Since both $C$'s firing and $E$'s firing are deviant, non-inertial events, the account would tell us that $C$'s firing caused $E$ to fire. But note that, in this model, the exogenous variable $A$ is inessential. So **Exogenous Reduction** tells us that it may be removed, leaving behind the model $\mathcal{M}_{10}^{-A}$,

$$E \coloneqq B \vee D$$
$$B \coloneqq \neg F$$
$$D \coloneqq F$$
$$F \coloneqq \neg C$$




And in this model, the interpolated endogenous variable $B$ is inessential. So **Endogeous Reduction** tells us that it may be removed, leaving behind the model $\mathcal{M}_{10}^{-A,-B}$,

$$E \coloneqq \neg F \vee D$$
$$D \coloneqq F$$
$$F \coloneqq \neg C$$




But in this model, $C \rightarrow F \rightarrow E$ would not count as a causal path. For $E = 1$ does not causally depend upon $F = 0$, since $F = 0$ is a default variable value, and $E = 1$ does not globally depend upon $F = 0$. Clause (c) in Causal Path thereby guarantees that productive causation is a model-invariant relation. (Note, however, that $E = 0$ does *locally* depend upon $F = 0$; so, had $F = 0$ been a deviant variable value, $E = 0$ would have causally depended upon $F = 0$, and $C \rightarrow F \rightarrow E$ would have been a causal path in $\mathcal{M}_{10}^{-A,-B}$.)

Productive causation is a somewhat complicated relation, but its complexities follow from a natural theory of production, together with the demand that our judgments not vary as we incorporate additional variables lying along, or leading into, the path from cause to effect. This helps to explain why a concept like productive causation is one worth having in the first place. *Inter alia*, our concept of causation earns its keep in our practices of blaming, praising, and explaining.[40]

[40] On causation's role in our practice of assigning moral responsibility, see Hart & Honoré (1985), Sartorio (2007, 2016), Moore (2009), and Schaffer (2012b).

In the prototypical case, we blame others for the bad consequences of their wrong actions and praise them for the good consequences of their supererogatory actions. Assume that we conceptualize especially good or bad states (and wrong or supererogatory acts) as deviant departures from the norm,[41] and assume that attributing moral responsibility for some state is a matter of tracing the propagation of this deviancy from state to act. Then our practice of attributing responsibility requires a concept of production. In the prototypical case of explanation, we wish to understand why some abnormal, non-inertial event took place. One way to understand why an abnormal event took place is to discover a source from which the abnormal behavior emanates. This, too, requires a notion of production.

Suppose we trace out a productive path from some bad consequence back to a wrong act; or suppose we trace out a productive path from some unexpected phenomenenon back to some deviant source. Upon closer inspection, we may find that what we once thought was production is not in fact the *uninterrupted* propagation of deviant states—perhaps, somewhere along the path from cause to effect, we discover a case of double prevention. Our practices of blaming and explaining need not be sensitive to such discoveries.[42] So we have reason to want a notion which is like production, but which ignores irrelevant details like whether the propagation of deviancy is achieved through double prevention. That is, we have reason to want a notion which is like production, but whose verdicts do not change as we attend to additional factors lying along, or feeding into, the path from cause to effect—a notion like production, but model-invariant. And productive causation is just such a notion.

Is productive causation just causation? Are all causes productive causes? If so, then counterfactual dependence does not suffice for causation. In many cases of omission (figure 4) and prevention (figure 5)—namely, those where the omitted or prevented events are deviant, or their omissions or preventions default—we would have dependence without causation. Interestingly, even with deviant, rather than default, events, there can be dependence without productive causation. First, consider the case of 'switching' shown in figure 16a. There, if $C$ fires, then any signal from $S$ will travel down to $D$. And if $C$ doesn't fire, then any signal from $S$ will travel up to $U$. The neurons $U$ and $D$ are both *dull*—they require two stimulatory signals in order to fire. Thus, $S$ functions as a switch, directing the signal from $F$ either upwards to $U$ or downwards to $D$. If $C$ fires, then it will flip the switch down, and $F$ will be a productive cause of $E$'s firing, and $A$ will not. Whereas, if $C$ doesn't fire, the switch will be up, and $A$ (as well as $F$) will be a productive cause of $E$'s firing. Assuming that $S$'s being in the up position is no more deviant than its being in the down position, $C$'s firing is a productive cause

[41] *Cf.* Hitchcock & Knobe (2009), Kahneman & Miller (1986).

[42] *Cf.* Schaffer (2012b).

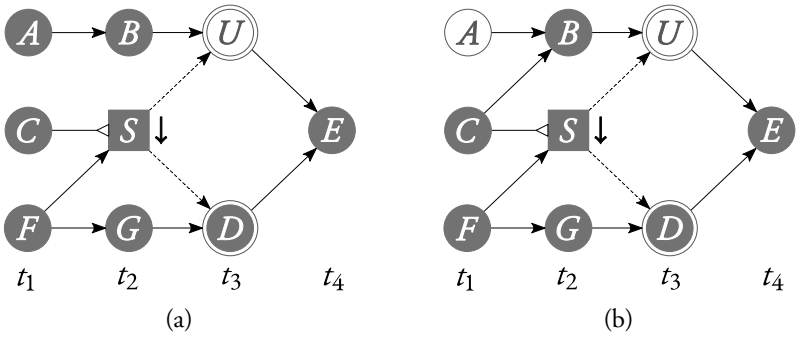



FIGURE 16: $S$ will fire iff $F$ fires. If $C$ additionally fires, then any signal from $S$ will travel downwards to $D$. If $F$ fires and $C$ doesn't, then the signal from $S$ will travel upwards to $U$. $U$ and $D$ are both *dull* neurons—they require two stimulatory signals in order to fire.

of $D$'s firing, but not of $E$'s firing.

Contrast figure 16a with figure 16b. Figure 16b is just like figure 16a, except that, in 16b, it is not $A$, but rather $C$, which initiates the signal traveling along the upper path through $B$. If $S$ had fired upwards, then $C$ would have been a productive cause of $E$, via the path $C \rightarrow B \rightarrow U \rightarrow E$. However, at the same time that $C$ makes $B$ fire, it also flips the switch so that the signal from $F$ travels downward to $D$. $F$'s firing is a productive cause of $E$'s firing in both figures 16a and 16b. $C$'s firing is not a productive cause of $E$'s firing in either figure 16a or 16b. However, in figure 16b, $E$'s firing globally depends upon $C$'s firing. So figure 16b provides a case of global dependence between deviant, rather than default, events which is not an instance of productive causation.

Does $C$'s firing cause $E$ to fire in figure 16b? I can see arguments on both sides. In figure 16a, $C$'s firing did not cause $E$ to fire by making $S$ firing downwards, and by making $S$ fire downwards, $C$'s firing robbed $A$ of its causal status. So, in figure 16b, when $C$ plays the same role in making $S$ fire downwards, and it plays the same role as $A$ in making $B$ fire, surely $C$ should likewise rob *itself* of its causal status by making $S$ fire downwards. On the other hand, if there were not a stimulatory connection between $C$ and $B$, so that $B$ hadn't fired, then $C$ would have been a cause (and a productive cause) of $E$'s firing. But the presence of the stimulatory connection between $C$ and $B$ doesn't make any difference with respect to what $C$ accomplishes along the path $C \rightarrow S \rightarrow D \rightarrow E$. So $C$'s firing should be a cause of $E$'s firing even when the connection between $C$ and $B$ is present.[43]

[43] An objection to this second argument: note that, in figure 16a, whether $C$'s firing is a cause of $E$'s firing depends upon whether $B$ fires. If it does, then $C$'s firing is not a cause of $E$'s firing. If not, then it is. In figure 16a, we should not treat the firing of $B$ as irrelevant to the question of whether $C$ caused $E$ to fire along the path $C \rightarrow S \rightarrow D \rightarrow E$. And so we likewise shouldn't

If productive causation is just causation, this explains some otherwise puzzling features of our causal thought and talk. To borrow an example from McGrath (2005), suppose that 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.[44] If we suppose that breaking a promise is deviant, and keeping it default, then Bob's failure to water the plant is a productive cause of its death. Likewise, if we suppose that Carlos's failure to water is default, then Carlos's failure to water is not a productive cause of its death.

There is some intuitive appeal to the idea that causation is just productive causation. There is also theoretical appeal to the idea that there are causes which are not productive causes. For it is natural to think that longer causal paths are built up out of shorter ones—that, if $c$ causes a distal $e$ by way of its consequences for the proximate $d$, then $c$ is also a cause of $d$, and $d$ is, in turn, a cause of $e$.[45] For instance, it is natural to think, in the case of double prevention from figure 13 that, if $C$'s firing caused $E$ to fire via the causal path $C \rightarrow D \rightarrow E$, then $C$'s firing should be a cause of $D$'s dormancy, and $D$'s dormancy, in turn, a cause of $E$'s firing. So it is natural to think that, once we recognize double prevention as a species of causation, we should likewise recognize both omission and prevention as species of causation as well. If we are moved by the thought that longer causal paths should be built up out of shorter ones, then we are led to the theory that causation is a hybrid of productive causation and counterfactual dependence—that $C = c$ caused $E = e$ iff either $C = c$ is a productive cause of $E = e$ or $E = e$ globally depends upon $C = c$. Both productive causation and global dependence are model-invariant relations, so this hybrid relation will also be model-invariant.

Alternatively, we could let in only dependence between deviant events. We could let in cases of prevention but not omission, omission but not prevention, or omission and prevention both, but not cases of omissive prevention (*e.g.*, I kept you in good health by not poisoning your food—you're welcome). Any of these accounts would be model-invariant. The amount of dependence we let into this theory is a free parameter which does not affect its verdicts in the more central cases of production or productive causation, and does not affect whether the theory is model-invariant.

---

treat the firing of $B$ as irrelevant to the question of whether $C$ caused $E$ to fire in figure 16b.

[44] See also the pen case in Hitchcock & Knobe (2009).

[45] See, *e.g.*, Price (1992).

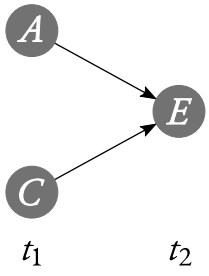



Figure 17: *Symmetric Overdetermination*

## 7 Symmetric Overdetermination

A simple case of symmetric overdetermination is shown in figure 17. Either $A$ 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 significant $A$'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 a case with a similar structure, consider Pay Raise.

> Pay Raise
> Franny, Sammy, and Tammy vote on a proposal to raise legislators' salaries. The proposal requires two out of three votes in order to pass. All three vote for the proposal, and it 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 both of these cases, there is an effect—$E$'s firing, the proposal's passing—which 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 are individually causes of the effect. $A$ and $C$ both individually caused $E$ to fire; and Franny, Sammy, and Tammy each individually caused the proposal to pass. At the same time, there is some appeal to the idea that Tammy's 'yea' vote didn't *all by itself* cause the proposal to pass. Perhaps she was a *part* of a cause—perhaps she *contributed* to the passing of the proposal—but, we might think, she did not cause it to pass all by herself.

Mackie (1965)[46] and Lewis (1986)[47] were both happy with the judgment that

[46] "...we would ordinarily hesitate to say, of either bullet, that it caused the man's death, or of either the lightning or the cigarette butt that it caused the fire...Our ordinary concept of cause does not deal clearly with cases of this sort." (Mackie, 1965, p. 251).

[47] "Such cases [of symmetric overdetermination] can be left as spoils to the victor, in D. M. Armstrong's phrase. We can reasonably accept as true whatever answer comes from the analysis that

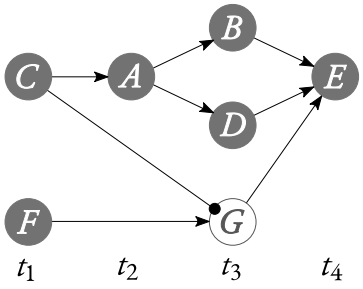



Figure 18

Tammy's 'yea' vote did not, all by itself, cause the proposal to pass in Pay Raise, and that $C$'s firing did not, all by itself, cause $E$ to fire in figure 17. According to both, in cases of symmetric overdetermination, intuition is split and a theory of causation could reasonably answer with either verdict. I agree with Mackie and Lewis (though this puts me in the minority of contemporary philosophers working on causation).[48]

An adequate theory of causation needn't say 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 causes $E$ to fire, then they must do so *jointly*. I will formally represent $A$ and $C$'s jointly causing $E$ to fire by allowing not just individual variable values, but also *vectors* of variable values, to be causes. In the canonical causal model $\mathcal{M}_{17}$, to say that $A$'s firing and $C$'s firing jointly caused $E$ to fire is to say that $(A, C) = (1, 1)$ caused $E = 1$.[49]

Once we admit vectors of variable values as causes, our theory must be generalized. To begin with, we should generalize the notion of a *causal path*. Consider the neuron network in figure 18. In the canonical causal model $\mathcal{M}_{18}$, $E = 1$ depends (both locally and globally) upon $(B, D) = (1, 1)$, rather than $(0, 0)$. However, $E = 1$ does not depend (either globally or locally) upon $C = 1$. Neither is there a causal path leading from $C$ to $E$. $C \rightarrow A \rightarrow B \rightarrow E$ is not a causal path, since $E = 1$ does not causally depend upon $B = 1$. And $C \rightarrow A \rightarrow D \rightarrow E$ is not a causal path, since $E = 1$ does not causally depend upon $D = 1$. But $C$'s firing caused $E$ to fire in figure 18.[50]

---

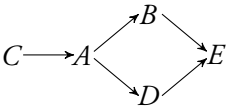
does best on the clearer cases." (Lewis, 1986, p. 194)

[48] Halpern & Pearl, Hitchcock, Woodward, and Weslake, *inter alia*, take it as a desideratum of a theory of causation that it say that Tammy caused the proposal to pass all by herself, and that $C$'s firing caused $E$ to fire all by itself.

[49] '$(A, C)$' is a vector whose first component is the variable $A$ and whose second component is the variable $C$; '$(1, 1)$' is a vector whose first and second components are the value 1. '$(A, C) = (1, 1)$' thus says that $A = 1$ and $C = 1$.

[50] As an aside, the reader may have wondered why, in our definition of causal dependence in §5, we did not allow deviant, rather than default, variable values related by global, but not local, dependence to be thereby related by causal dependence. Figure 18 illustrates why. If we allowed

Let us begin by generalizing the notion of a directed path. A directed path, $\mathcal{P}$, from $C$ to $E$ can be understood as a collection of directed edges between variables generated by the following procedure: begin with $E$, and select exactly one of its causal parents, $P$, to be its $\mathcal{P}$-parent. Then, include the directed edge between $E$ and $P$, $P \to E$, in $\mathcal{P}$. Next, select exactly one of $P$'s causal parents to be its $\mathcal{P}$-parent, and proceed in this manner until you reach $C$. We can then define a directed *forking* path, $\mathcal{F}$, from the *vector* of variables $\mathbf{C}$ to $E$, as a collection of directed edges generated by the following procedure: begin with $E$, and select some of its causal parents, $P_1, P_2, \ldots, P_N$ (you needn't choose just one) to be its $\mathcal{F}$-parents. Include each of the directed edges between the $P_i$ and $E$, $P_i \to E$, in $\mathcal{F}$. Next, for each of the $P_i$, select some of their causal parents to be their $\mathcal{F}$-parents, and proceed in this manner until you have reached all and only the variables in $\mathbf{C}$. For instance, in $\mathcal{M}_{18}$, $B \to E \leftarrow D$ is a directed forking path from $(B, D)$ to $E$, $C \to A \to D \to E \leftarrow G \leftarrow F$ is a directed forking path from $(C, F)$ to $E$, and




is a directed forking path from $C$ to $E$.

$C$'s firing caused $E$ to fire in figure 18 because this directed forking path is causal. What it is for a directed *forking* path to be causal is just what it is for a directed path to be causal, *mutatis mutandis*. First, some terminology: if there is a directed edge between $U$ and $V$ in $\mathcal{F}$, then $U$ is one of $V$'s $\mathcal{F}$-parents.[51] And if there is a sequence of directed edges in $\mathcal{F}$ leading from $U$ to $V$, then $V$ is an $\mathcal{F}$-descendant of $U$. If a directed forking path is to be causal, then, in the first place, there must be an assignment of contrasts such that every variable value along the path, rather than its contrast, causally depends upon the vector of its $\mathcal{F}$-parents' values, rather than their contrasts.

Suppose that we are given a directed forking path $\mathcal{F}$, from $\mathbf{C}$ to $E$, and along this path lie two variables, $D$ and $R$, such that $R$ is an $\mathcal{F}$-descendant of $D$. If there is a *separate* directed path from $D$ to $R$, $\mathcal{O} : D \to O_1 \to O_2 \to \cdots \to O_M \to R$ such that none of the directed edges from $\mathcal{O}$ appear in $\mathcal{P}$, then say that $D$ is a *departure* variable, and $R$ is one of its *return* variables (relative to the forking path $\mathcal{F}$). For instance, in $\mathcal{M}_{18}$, relative to the directed forking path leading from $C$ to $E$ (through $A, B$, and $D$), $C$ is a departure variable, and $E$ is its return.

---

global dependence between deviant, rather than default, variable values to suffice for causal dependence, then, in $\mathcal{M}_{18}^{-B}$, $C \to A \to E$ would be a causal path. So the account from §6 would not have been model-invariant.

[51] Note: there is a distinction between a variable's $\mathcal{F}$-parents and its causal parents on $\mathcal{F}$. In $\mathcal{M}_6$, $\mathcal{F} : C \to B \to E$ is a directed forking path, and $C$ is a causal parent of $E$ on $\mathcal{F}$; but it is not one of its $\mathcal{F}$-parents. Contrast $\mathcal{F}$ with $\mathcal{F}^\star : C \to B \to E \leftarrow C$. $C$ *is* an $\mathcal{F}^\star$-parent of $E$.

Take some forking path $\mathcal{F}$, with a departure-return variable pair, $< D, R >$. If $\mathcal{F}$ is to be causal, then $D$ must be assigned some contrast value—call it '$d^*$'. Let $\mathbf{I}$ be the intermediate variables which are $\mathcal{F}$-descendants of $D$ and $\mathcal{F}$-ancestors of $R$ (excluding $D$ and $R$ themselves), and let their designated contrasts be $\mathbf{i}^*$. Then, we will say that $< D, R >$ is an *active* departure-return variable pair (relative to the forking path $\mathcal{F}$ and an assignment of contrasts to the variables on $\mathcal{F}$) if there is some (perhaps empty) collection of variables on the forking path, $\mathbf{F}$, with designated contrasts $\mathbf{f}^*$, such that some parent of $R$ which is not on the path $\mathcal{F}$ takes on a different value in the counterfactual model $\mathcal{M}[\mathbf{F} \rightarrow \mathbf{f}^*, \mathbf{I} \rightarrow \mathbf{i}^*, D \rightarrow d^*]$ than it does in the model $\mathcal{M}[\mathbf{F} \rightarrow \mathbf{f}^*, \mathbf{I} \rightarrow \mathbf{i}^*]$. If $< D, R >$ is an active departure-return variable pair, then, when some of the variables on the path, including any intermediate between $D$ and $R$, are held fixed at their contrast values, wiggling $D$ wiggles some non-$\mathcal{F}$ parent of $R$. In that case, $D$ potentially affects $R$ both along $\mathcal{F}$ and along some other path or paths. It could be that, what $D$ gives along $\mathcal{F}$, it takes away along some other path. If $D$ gives a deviant, rather than a default, value, then this will make no difference with respect to whether $\mathcal{F}$ is a forking causal path. But if $D$ and $R$ are active, and $D$ does not give a deviant value to $R$, then $\mathcal{F}$ is not causal.[52]

Then, we may say that a directed forking path is causal only if (a) each variable value, rather than its contrast, causally depends upon its parents on the path, rather than their contrasts; (b) the initial and final variables on the path take on deviant, rather than default, values; and (c) every active departure-return variable pair takes on deviant, rather than default, values.

> Forking Causal Path (provisional)
> A directed forking path, $\mathcal{F}$, from $\mathbf{C}$ to $E$ is a *forking causal path* leading from $\mathbf{C}$ to $E$ iff there is an assignment of contrasts to the variables along $\mathcal{F}$ such that:
>
> (a) for each $V \notin \mathbf{C}$ along the path, $V$'s value, rather than its contrast, causally depends upon $V$'s $\mathcal{F}$-parents' values, rather than their contrasts.
>
> (b) The variables in $\mathbf{C} \cup (E)$ have deviant values and default contrasts.
>
> (c) For every active departure-return variable pair along $\mathcal{F}$, $< D, R >$, $D$ and $R$ have deviant values and default contrasts.

When we were only considering individual variables as causes, we said that, when variables take on deviant, rather than default, values, causal dependence is local dependence; otherwise, causal dependence is global dependence. When we

[52] Couldn't an active departure $D$ have a default value or deviant contrast, yet still not take away along other paths what it gives $R$ along $\mathcal{F}$? Yes, but in this case, the additional paths from $D$ to $R$ may simply be incorporated into $\mathcal{F}$.

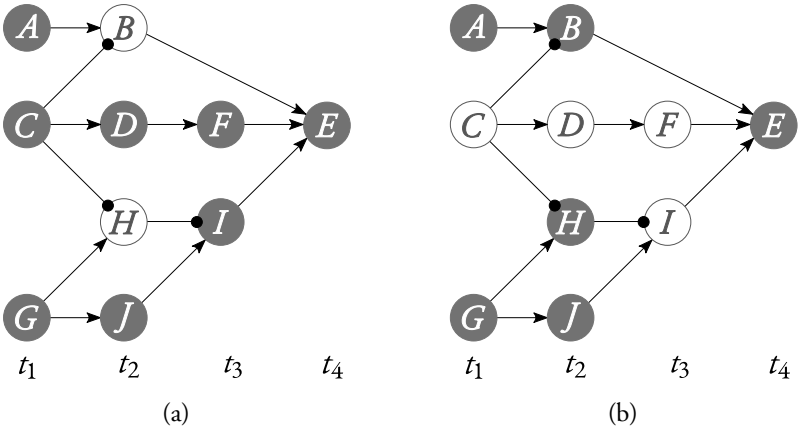
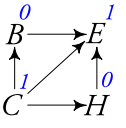


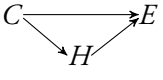

FIGURE 19: 19b shows what would have happened in 19a, had $C$ not fired.

turn our attention to vectors of variable values, it could be that one component of the vector is deviant, while another component of the vector is default. For instance, take the neuron network from figure 19a. There, $C$'s firing causes $E$ to fire. Had $C$ not fired, as in figure 19b, $A$ would have caused $E$ to fire by itself; but $C$ preempts the causal process beginning with $A$ and initiates its own causal process with $D$'s firing which runs to completion. The causal process running through $D$ and $F$ was sufficient, all by itself, to make $E$ fire; however, at the same time as $C$ initiates this causal process, it initiates another as well, by preventing $H$ from preventing $I$ from firing. So $C$ preempts $A$ and supplies more than enough to make $E$ fire. As the reader may verify for themselves, by successively removing inessential variables from the canonical model of figure 19a, we may arrive at the model $\mathcal{M}^{\star}_{19a}$,

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




In this model, $E = 1$ does not counterfactually depend (either locally or globally) upon $C = 1$ individually. There's no global dependence because, had $C$ not fired, $B$ would have, and $E$ would have fired all the same. There's no local dependence because, holding fixed that $H$ didn't fire, $E$ would have fired whether or not $C$ did. So $C \rightarrow E$ is not a causal path. We wish to say that $C = 1$ caused $E = 1$ by way of the forking causal path

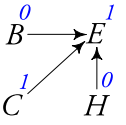



But $H$ takes on a default value, and $E = 1$ does not globally depend upon the vector of values $(C, H) = (1, 0)$, rather than $(0, 1)$. When dealing with a vector of

variable values like $(C, H) = (1, 0)$, where one variable is deviant and the other is default, how do we understand causal dependence?

Let us begin by generalizing the notion of a local model. Given a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with $E \in \mathcal{V}$, and some vector of variables $\mathbf{P} \subseteq \mathbf{PA}(E)$, let us define the $\mathbf{P}$-*local* model at $E$, $\mathcal{M}(E)_{\mathbf{p}}$, as the model that you get by the following procedure: for each $P \in \mathbf{P}$, replace $P$ with its actual value wherever it appears in any structural equation $\phi_V \in \mathcal{E}$ other than $\phi_E$. If $P$ was the only variable appearing in $\phi_V$, then we may go ahead and exogenize the variable $V$.[53] For instance, in the model of figure 19a from above, the $C$-local model at $E$ is

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




Then, given a vector of parents of $E$, $\mathbf{C} \subseteq \mathbf{PA}(E)$, with $\mathbf{P} \subseteq \mathbf{C}$, $E = e$, rather than $e^*$, $\mathbf{P}$-locally depends upon $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, if and only if $E = e$, rather than $e^*$, depends upon $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, in the $\mathbf{P}$-local model at $E$,

$$\mathcal{M}(E)_{\mathbf{p}} \models \mathbf{C} = \mathbf{c}^* \; \square\!\!\rightarrow E = e^*$$

For instance, in the model of figure 19a from above, $E = 1$, rather than 0, $C$-locally depends upon $(C, H) = (1, 0)$, rather than $(0, 1)$.

$$\mathcal{M}^{\star}_{19a}(E)_C \models (C, H) = (0, 1) \; \square\!\!\rightarrow E = 0$$

Note that $\mathbf{P}$-local dependence between $E$ and $\mathbf{C}$ is only defined when $\mathbf{P} \subseteq \mathbf{C} \subseteq \mathbf{PA}(E)$.

Finally, we can say that $E = e$, rather than $e^*$, *causally* depends upon a *vector* of its parent variables' values, $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, iff $E = e$, rather than $e^*$, $\mathbf{Dev}$-locally depends upon $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, where $\mathbf{Dev}$ is the subvector of $\mathbf{C}$ containing the variables in $\mathbf{C}$ which take on deviant values in $\mathbf{c}$ and default values in $\mathbf{c}^*$. For instance, in the model of figure 19a, $E = 1$, rather than 0, causally depends upon $(C, H) = (1, 0)$, rather than $(0, 1)$.

> Causal Dependence$^{\star}$
> For $\mathbf{C} \subseteq \mathbf{PA}(E)$, $E = e$, rather than $e^*$, *causally depends* upon $\mathbf{C} = \mathbf{c}$, rather than $\mathbf{c}^*$, iff
>
> $$\mathcal{M}(E)_{\mathbf{Dev}} \models \mathbf{C} = \mathbf{c}^* \; \square\!\!\rightarrow E = e^*$$
>
> where $\mathbf{Dev} \subseteq \mathbf{C}$ is the vector of variables which take on deviant values in $\mathbf{c}$ and default values in $\mathbf{c}^*$.

[53] To exogenize a variable $V \in \mathcal{V}$: move $V$ from $\mathcal{V}$ to $\mathcal{U}$, enrich the exogenous assignment $\mathbf{u}$ so that it assigns $V$ the value it takes on in the original model, and remove $V$'s structural equation $\phi_V$ from $\mathcal{E}$.

Note that, in the case where $\mathbf{C}$ is a 1-vector containing a single variable, $\mathbf{c}$ is a deviant value, and $\mathbf{c}^*$ is a default value, causal dependence is local dependence.[54] Note also that, in the case where $\mathbf{C}$ is a 1-vector containing a single variable and either $\mathbf{c}$ isn't deviant or $\mathbf{c}^*$ isn't default, causal dependence is global dependence. In mixed cases like the one above, testing for causal dependence involves making local counterfactual assumptions about deviant, rather than default, variable values and global counterfactual assumptions about other variable values. Thus, $E = 1$ causally depends upon $(C, H) = (1, 0)$, rather than $(0, 1)$, and $H = 0$, rather than 1, causally depends upon $C = 1$, rather than 0. So $C \rightarrow H \rightarrow E \leftarrow C$ is a forking causal path.

We are now in a position to formulate a notion of productive causation which covers joint as well as individual causation.

> Productive Causation$^\star$
> In a causal model $\mathcal{M}$, $\mathbf{C} = \mathbf{c}$ is a *productive cause* of $E = e$ if and only if there is a forking causal path, $\mathcal{F}$, leading from $\mathbf{C}$ to $E$, and there is no subpath of $\mathcal{F}$ leading from any proper subvector of $\mathbf{C}$ to $E$ which is itself a forking causal path.

I have included a minimality condition stipulating that a vector of variables $\mathbf{C}$ productively caused $E = e$ along a forking path $\mathcal{F}$ only if no subvector of those variables caused $E = e$ along a subpath of $\mathcal{F}$.[55] To appreciate why, consider the neuron network from figure 20a. (There, $E$ is a *dull* neuron. It will only fire if at least two of its parent neurons do.) In the canonical causal model $\mathcal{M}_{20a}$, there is a forking causal path leading from $(C, A)$ to $E$. For $E = 1$ causally depends upon $(B, D, G) = (1, 1, 0)$, rather than $(0, 0, 1)$; $B = 1$, rather than 0, causally depends upon $C = 1$, rather than 0; $D = 1$, rather than 0, causally depends upon $C = 1$, rather than 0; and $G = 0$, rather than 1, causally depends upon $A = 1$, rather than 0. But $A$ is not a joint cause of $E$'s firing, along with $C$. The minimality condition rules out spurious joint causes like these. For $C \rightarrow B \rightarrow E \leftarrow D \leftarrow C$ is a subpath of the forking causal path leading from $(A, C)$ to $E$, and this subpath is causal.

Why not simply say '$\mathbf{C}$ caused $E$ only if there is no forking causal path from any proper subvector of $\mathbf{C}$ to $E$'? Because then we could not say that, in figure 20b, $C$'s firing is a joint cause of $E$'s firing. For even though there is a forking causal path from $(A, C)$ to $E$, namely $C \rightarrow E \leftarrow A$, there is also a forking causal path from $A$ alone to $E$, namely $A \rightarrow C \rightarrow E \leftarrow A$. $A$ is a cause of $E$'s firing, all by itself. But, even so, $A$ and $C$ jointly cause $E$ to fire.[56] The proposed minimality

[54] Throughout, we should draw no distinction between a 1-dimensional vector and its component.

[55] A *subpath* of a forking path $\mathcal{F}$ is a subset of the directed edges in $\mathcal{F}$ which is itself a directed forking path.

[56] *Cf.* Rosenberg & Glymour (forthcoming).

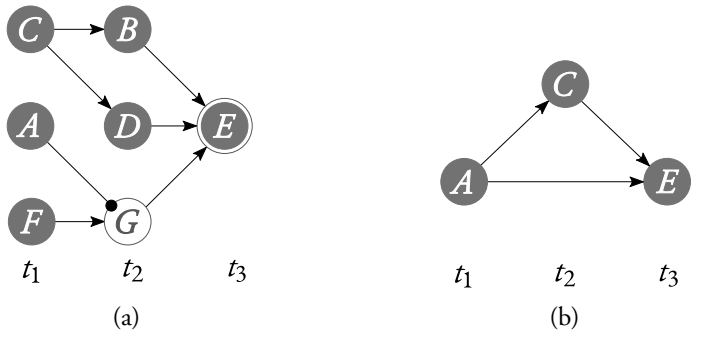



FIGURE 20: In figure 20a, $E$ is a *dull* neuron. It will only fire if at least two of its parent neurons fire.

condition must disagree. But, since $A \rightarrow C \rightarrow E \leftarrow A$ is not a subpath of $C \rightarrow E \leftarrow A$, our minimality condition allows us to agree.

Are joint causes causes simpliciter? In the case of symmetric overdetermination from figure 17, for instance, is $C$'s firing a cause of $E$'s firing? We could go either way. We could decide to say that the components of a vector cause are causes in their own right. Or we could decide to say that they are merely *parts* of a cause, and distinguish joint from individual causation. My own inclination is to say that neither $A$ nor $C$ individually caused $E$ to fire in figure 17, even though, together, they did; but if the reader balks at this, they should feel free to go the other way.

If we only consider 1-vectors as potential causes and don't allow causal paths to fork, PRODUCTIVE CAUSATION$^\star$ reduces to PRODUCTIVE CAUSATION from §6.[57] So the relation of productive causation defined here is a strict generalization of the one from §6. So generalized, productive causation is still a model-invariant relation. Suppose we have a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with $\mathbf{C} \subseteq \mathcal{U} \cup \mathcal{V}$, $E \in \mathcal{V}$, and some inessential $U \in \mathcal{U}$. If $\mathbf{C} = \mathbf{c}$ is a productive cause of $E = e$ in $\mathcal{M}$, then, so long as $U \notin \mathbf{C}$, $\mathbf{C} = \mathbf{c}$ will also be a productive cause of $E = e$ in $\mathcal{M}^{-U}$. Or suppose we have a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with $\mathbf{C} \subseteq \mathcal{U} \cup \mathcal{V}$, $E \in \mathcal{V}$, and some inessential $V \in \mathcal{V}$. If $\mathbf{C} = \mathbf{c}$ is a productive cause of $E = e$ in $\mathcal{M}$, then, so long as $V \notin \mathbf{C} \cup (E)$, $\mathbf{C} = \mathbf{c}$ will still be a productive cause of $E = e$ in $\mathcal{M}^{-V}$. (See **Proposition 2** in the appendix.)

Contrast joint productive causation with joint production. $\mathbf{C} = \mathbf{c}$ produces $E = e$ iff there is a productive forking path leading from $\mathbf{C}$ to $E$. A directed forking path from $\mathbf{C}$ to $E$, $\mathcal{F}$, is a productive forking path iff there is an assignment of contrasts to the variables along $\mathcal{F}$ such that (a) each variable $V \notin \mathbf{C}$ along $\mathcal{F}$ causally depends upon its $\mathcal{F}$-parents; and (b) every variable along $\mathcal{F}$ takes on a deviant, rather than a default, value. If $\mathbf{C} = \mathbf{c}$ produces $E = e$, then $\mathbf{C} = \mathbf{c}$ is a pro-

[57] Again, we shouldn't distinguish a variable value $C = c$ from a 1-vector variable value $(C) = (c)$.

ductive cause of $E = e$, though the converse is false. This notion of production is model-variant. Every bit of the additional weakness in Productive Causation$^\star$ is required in order for joint productive causation to be model-invariant. So just as in §6, we may view joint causation as a natural model-invariant weakening of a relation of production.

As before, we could say that causation is productive causation, or we could say that it is a hybrid of productive causation and global dependence—that $\mathbf{C} = \mathbf{c}$ caused $E = e$ iff either $\mathbf{C} = \mathbf{c}$ is a productive cause of $E = e$ or $E = e$ globally depends upon $\mathbf{C} = \mathbf{c}$ (and no subvector thereof). Both productive causation and global dependence are model-invariant relations, so the hybrid relation will also be model-invariant.

## A Technicalities

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

**Lemma 1.** *Given a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with $C \in \mathcal{U} \cup \mathcal{V}$, $E, V \in \mathcal{V}$, and $V \neq C, E$, if $V$ is inessential, then there is a causal path from $C$ to $E$ in $\mathcal{M}$ if and only if there is a causal path from $C$ to $E$ in $\mathcal{M}^{-V}$.*

*Proof.* We first establish the 'only if' direction. Assume that there is such a causal path, $\mathcal{P}$, in $\mathcal{M}$. Since $V$ is inessential, it has a single parent, $Pa$, and at most a single child, $Ch$ (and $Pa$ is not a parent of $Ch$). Let their actual values in $\mathcal{M}$ be $v$, $pa$, and $ch$, respectively. There are two possibilities: either (A) $V$ does not lie on $\mathcal{P}$; or (B) it does. In case (A), removing $V$ may introduce new causal dependence relationships between $Pa$ and $Ch$, but it will not alter any causal dependence relations between any of the variables on $\mathcal{P}$ and their $\mathcal{P}$-parents. Since, in $\mathcal{M}$, each variable along $\mathcal{P}$, rather than its contrast, causally depends upon its $\mathcal{P}$-parent's value, rather than its contrasts, in $\mathcal{M}^{-V}$, each variable along $\mathcal{P}$, rather than its contrast, will causally depend upon its $\mathcal{P}$-parent's value, rather than its contrasts. For any departure-return pair variable, $< D, R >$, along $\mathcal{P}$, removing $V$ will not affect $< D, R >$ is active, nor whether $(r, r^*)$ and $(d, d^*)$ are deviant, rather than default. So, in case (A), $\mathcal{P}$ will still be a causal path in $\mathcal{M}^{-V}$. In case (B), $V$ lies on $\mathcal{P}$. Then, $Pa$ and $Ch$ must lie on $\mathcal{P}$ as well. Let $\mathbf{P}$ be $Ch$'s parents other than $V$ (if such there be); and let their actual values be $\mathbf{p}$. Then, in $\mathcal{M}$, there are some $v^*$, $pa^*$, and $ch^*$ such that $(ch, ch^*)$ causally depends upon $(v, v^*)$ and $(v, v^*)$ causally depends upon $(pa, pa^*)$. Since $Pa$ is $V$'s only parent, we can conclude that

$$\phi_V(pa^*) = v^* \tag{1}$$

Since $Ch$ is $V$'s only causal child, there is no difference between local and global dependence; so, since $(ch, ch^*)$ causally depends upon $(v, v^*)$, we can conclude that

$$\phi_{Ch}(v^*, \mathbf{p}) = ch^* \tag{2}$$

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

$$Ch := \phi_{Ch}(\phi_V(Pa), \mathbf{P})$$

In $\mathcal{M}$, either (B-1) $< Pa, Ch >$ is not an active departure-return variable pair, relative to $\mathcal{P}$ and the relevant assignment of contrasts, or (B-2) it is. In case (B-1), note that (3) follows from (1) and (2).

$$\phi_{Ch}(\phi_V(pa^*), \mathbf{p}) = ch^* \tag{3}$$

Since $< Pa, Ch >$ is not an active departure-return variable pair,

$$\mathcal{M}^{-V}[Pa \rightarrow pa^*] \models \mathbf{P} = \mathbf{p} \tag{4}$$

It then follows from (3) and (4) that, in $\mathcal{M}^{-V}$, $(ch, ch^*)$ causally depends upon $(pa, pa^*)$.

So the truncated $\mathcal{P} - V$[58] will be a causal path in $\mathcal{M}^{-V}$. In case (B-2), $< Pa, Ch >$ is an active departure-return variable pair, relative to $\mathcal{P}$ and the relevant assignment of contrasts, in $\mathcal{M}$. Since $\mathcal{P}$ is a causal path in $\mathcal{M}$, $(pa, pa^*)$ and $(ch, ch^*)$ are both deviant, rather than default. It then follows from (3) that $(ch, ch^*)$ causally depends upon $(pa, pa^*)$. So the truncated $\mathcal{P} - V$ will be a causal path in $\mathcal{M}^{-V}$. So, in either case under (B), there will be a causal path from $C$ to $E$ in $\mathcal{M}^{-V}$.

To establish the 'if' direction, suppose that there is a causal path, $\mathcal{P}$, from $C$ to $E$ in $\mathcal{M}^{-V}$. $\mathcal{P}$ either (A) includes the directed edge $Pa \rightarrow Ch$ or (B) doesn't. If (A), then either (A-1) $< Pa, Ch >$ is an active departure-return variable pair relative to $\mathcal{P}$ and the relevant assignment of contrasts, or (A-2) it isn't. If (A-1), then there must be some $pa^*$ and $ch^*$ such that $(pa, pa^*)$ and $(ch, ch^*)$ are both deviant, rather than default, and $(ch, ch^*)$ locally depends upon $(pa, pa^*)$. So

$$\phi_{Ch}(\phi_V(pa^*), \mathbf{p}) = ch^* \tag{5}$$

(**P** are the parents of $Ch$ other than $Pa$, and **p** are their values.) Let $v^*$ be the value of $V$ such that $v^* = \phi_V(pa^*)$. Then, since $Ch$ is $V$'s only causal child, it follows from (5) that in $\mathcal{M}$, $(ch, ch^*)$ will causally depend upon $(v, v^*)$. Since $Pa$ is $V$'s only parent in $\mathcal{M}$, it also follows that $(v, v^*)$ causally depends upon $(pa, pa^*)$. So the elongated $\mathcal{P} + V$[59] will be a causal path in $\mathcal{M}$. If (A-2), then either (A-2-a) $(pa, pa^*)$ and $(ch, ch^*)$ are both deviant, rather than default, or (A-2-b) not. If (A-2-a), then (6) must hold.

$$\mathcal{M}^{-V}(Ch)[Pa \rightarrow pa^*] \models \mathbf{P} = \mathbf{p} \tag{6}$$

Again, let $v^*$ be the value of $V$ such that $v^* = \phi_V(pa^*)$. Then, since $Ch$ is $V$'s only causal child, it follows from (5) and (6) that in $\mathcal{M}$, $(ch, ch^*)$ will causally depend upon $(v, v^*)$. And $(v, v^*)$ will causally depend upon $(pa, pa^*)$. So the elongated $\mathcal{F} + V$ will be a causal path in $\mathcal{M}$. If (A-2-b), then $< Pa, Ch >$ could not be an active departure-return variable pair, so (7) must hold.

$$\mathcal{M}^{-V}[Pa \rightarrow pa^*] \models \mathbf{P} = \mathbf{p} \tag{7}$$

Again, let $v^*$ be the value of $V$ such that $v^* = \phi_V(pa^*)$. Then, since $Ch$ is $V$'s only causal child, it follows from (5) and (7) that in $\mathcal{M}$, $(ch, ch^*)$ will causally depend upon $(v, v^*)$. And $(v, v^*)$ will causally depend upon $(pa, pa^*)$. So the elongated $\mathcal{P} + V$ will be a causal path in $\mathcal{M}$. If (B), then $\mathcal{P}$ will also be a causal path in $\mathcal{M}$, since including the interpolated variable $V$ will not alter any of the causal dependence relationships amongst any of the variables other than $Pa$ and $Ch$. □

**Proposition 1.** Productive Causation *defines a model-invariant relation. That is: (a) given a causal model* $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$*, with* $U \in \mathcal{U}$*,* $C \in \mathcal{U} \cup \mathcal{V}$*,* $E \in \mathcal{V}$*, and* $U \neq C$*,* $C = c$ *is a productive cause of* $E = e$ *in* $\mathcal{M}$ *iff* $C = c$ *is a productive cause of* $E = e$ *in* $\mathcal{M}^{-U}$*. And (b) given a causal model* $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$*, with* $C \in \mathcal{U} \cup \mathcal{V}$*,* $E, V \in \mathcal{V}$*, and*

---

[58] This is the directed path $\mathcal{P}$, minus the directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$, and plus the new directed edge $Pa \rightarrow Ch$.

[59] This is the directed path $\mathcal{P}$, minus the directed edge $Pa \rightarrow Ch$, and plus the new directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$.

*$V \neq C, E$, if $V$ is inessential, then $C = c$ is a productive cause of $E = e$ in $\mathcal{M}$ iff $C = c$ is a productive cause of $E = e$ in $\mathcal{M}^{-V}$.*

*Proof.* For part (a): if $C = c$ is a productive cause of $E = e$ in $\mathcal{M}$, then, in $\mathcal{M}$, there is a causal path from $C$ to $E$. The exogenous $U \in \mathcal{U}$ will not be on this causal path, so removing it will not affect any of the causal dependence relationships between any of the variables on the path. Nor will it affect whether any departure and return variables are active or deviant. So there will be a causal path in $\mathcal{M}^{-U}$. If $C = c$ was not a productive cause of $E = e$ in $\mathcal{M}$, then there was no causal path from $C$ to $E$. The exogenous $U \in \mathcal{U}$ is not on any directed path from $C$ to $E$, and removing it will not affect any of the causal dependence relationships between any of the variables on any path from $C$ to $E$, nor whether any departure and return variables are active or deviant. So removing $U$ will not create any new causal paths. So $C = c$ will not be a productive cause of $E = e$ in $\mathcal{M}^{-U}$. Part (b) follows immediately from **Lemma 1**. □

**Lemma 2.** *Given a causal model $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$, with $\mathbf{C} \subseteq \mathcal{U} \cup \mathcal{V}$, $E, V \in \mathcal{V}$, and $V \notin \mathbf{C} \cup (E)$, if $V$ is inessential, then there is a forking causal path from $\mathbf{C}$ to $E$ in $\mathcal{M}$ if and only if there is a forking causal path from $\mathbf{C}$ to $E$ in $\mathcal{M}^{-V}$.*

*Proof.* We first establish the 'only if' direction. Suppose that there is such a forking causal path, $\mathcal{F}$, in $\mathcal{M}$. Since $V$ is inessential, it has a single parent, $Pa$, and at most a single child, $Ch$ (and $Pa$ is not a parent of $Ch$). Let their actual values in $\mathcal{M}$ be $v$, $pa$, and $ch$, respectively. There are two possibilities: either (A) $V$ does not lie on $\mathcal{F}$; or (B) $V$ does lie on $\mathcal{F}$. In case (A), removing $V$ may introduce new causal dependence relationships between $Pa$ and $Ch$, but it will not alter any causal dependence relations between any of the variables on $\mathcal{F}$ and their $\mathcal{F}$-parents. Since, in $\mathcal{M}$, each variable along $\mathcal{F}$, rather than its contrast, causally depends upon its $\mathcal{F}$-parents's values, rather than their contrasts, in $\mathcal{M}^{-V}$, each variable along $\mathcal{F}$, rather than its contrast, will still causally depend upon its $\mathcal{F}$-parents, rather than their contrasts. For any departure-return variable pair $< D, R >$ along $\mathcal{F}$, removing $V$ will not affect whether $< D, R >$ is active, nor whether $(r, r^*)$ and $(d, d^*)$ are deviant, rather than default. So, in case (A), $\mathcal{F}$ will still be a forking causal path in $\mathcal{M}^{-V}$. In case (B), $V$ lies on $\mathcal{F}$. Then, $Pa$ and $Ch$ must lie on $\mathcal{F}$ as well. Let $\mathbf{P}_{\mathcal{F}}$ be $Ch$'s parents other than $V$ that lie on the path $\mathcal{F}$ (if such there be); let their actual values be $\mathbf{p}_{\mathcal{F}}$, and their designated contrasts, $\mathbf{p}^*_{\mathcal{F}}$. Similarly, let $\mathbf{P}_{\overline{\mathcal{F}}}$ be $Ch$'s parents that don't lie on the path $\mathcal{F}$ (if such there be). Then, in $\mathcal{M}$, there are some $v^*$, $pa^*$, and $ch^*$ such that $(ch, ch^*)$ causally depends upon $(\mathbf{p}_{\mathcal{F}} \cup (v), \mathbf{p}^*_{\mathcal{F}} \cup (v^*))$ and $(v, v^*)$ causally depends upon $(pa, pa^*)$. Since $Pa$ is $V$'s only parent, we can conclude that

$$\phi_V(pa^*) = v^* \tag{8}$$

And since $(ch, ch^*)$ causally depends upon $(\mathbf{p}_{\mathcal{F}} \cup (v), \mathbf{p}^*_{\mathcal{F}} \cup (v^*))$, we can conclude that

$$\phi_{Ch}(v^*, \mathbf{p}^*_{\mathcal{F}}, \mathbf{p}^\star_{\overline{\mathcal{F}}}) = ch^* \tag{9}$$

(where $\mathbf{p}^\star_{\overline{\mathcal{F}}}$ are the values $\mathbf{P}_{\overline{\mathcal{F}}}$ take on in the model $\mathcal{M}(Ch)_{\mathbf{Dev}}[V \rightarrow v^*, \mathbf{P}_{\mathcal{F}} \rightarrow \mathbf{p}^*_{\mathcal{F}}]$, and **Dev** are the variables in $\mathbf{P}_{\mathcal{F}} \cup (V)$ which take on deviant, rather than default, values.) By the construction of $\mathcal{M}^{-V}$, it contains the structural equation

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

In $\mathcal{M}$, either (B-1) $< Pa, Ch >$ is not an active departure-return variable pair, relative to $\mathcal{F}$ and the relevant contrasts, or (B-2) it is. In case (B-1), note that (10) follows from (8) and (9).

$$\phi_{Ch}(\phi_V(pa^*), \mathbf{p}^*_{\mathcal{F}}, \mathbf{p}^\star_{\overline{\mathcal{F}}}) = ch^* \tag{10}$$

Since $< Pa, Ch >$ is not an active departure-return variable pair,

$$\mathcal{M}^{-V}(Ch)_{\mathbf{Dev}'}[Pa \rightarrow pa^*, \mathbf{P}_{\mathcal{F}} \rightarrow \mathbf{p}^*_{\mathcal{F}}] \models \mathbf{P}_{\overline{\mathcal{F}}} = \mathbf{p}^\star_{\overline{\mathcal{F}}} \tag{11}$$

(where $\mathbf{Dev}'$ are those variables in $\mathbf{P}_{\mathcal{F}} \cup (Pa)$ which take on deviant, rather than default, values). It then follows from (10) and (11) that, in $\mathcal{M}^{-V}$, $(ch, ch^*)$ causally depends upon $(\mathbf{p}_{\mathcal{F}} \cup (pa), \mathbf{p}^*_{\mathcal{F}} \cup (pa^*))$. So the truncated $\mathcal{F} - V$[60] will be a forking causal path in $\mathcal{M}^{-V}$. In case (B-2), $Pa$ and $Ch$ are an active departure-return variable pair, relative to $\mathcal{F}$ and the relevant contrasts, in $\mathcal{M}$. Since $\mathcal{F}$ is a forking causal path in $\mathcal{M}$, $(pa, pa^*)$ and $(ch, ch^*)$ are both deviant, rather than default. It then follows from (10) that $(ch, ch^*)$ causally depends upon $(\mathbf{p}_{\mathcal{F}} \cup (pa), \mathbf{p}^*_{\mathcal{F}} \cup (pa^*))$. So the truncated $\mathcal{F} - V$ will be a forking causal path in $\mathcal{M}^{-V}$. So, in either case under (B), there will be a forking causal path in $\mathcal{M}^{-V}$.

To establish the 'if' direction, suppose that there is a forking causal path, $\mathcal{F}$, from $\mathbf{C}$ to $E$ in $\mathcal{M}^{-V}$. $\mathcal{F}$ either (A) includes the directed edge $Pa \rightarrow Ch$ or (B) doesn't. If (A), then there must be some $pa^*$, $ch^*$, and $\mathbf{p}^*_{\mathcal{F}}$ such that $(ch, ch^*)$ causally depends upon $(\mathbf{p}_{\mathcal{F}} \cup (pa), \mathbf{p}^*_{\mathcal{F}} \cup (pa^*))$. ($\mathbf{P}_{\mathcal{F}}$ are $Ch$'s $\mathcal{F}$-parents, other than $Pa$, if such there be.) So

$$\phi_{Ch}(\phi_V(pa^*), \mathbf{p}^*_{\mathcal{F}}, \mathbf{p}^\star_{\overline{\mathcal{F}}}) = ch^* \tag{12}$$

($\mathbf{P}_{\overline{\mathcal{F}}}$ are the parents of $Ch$ which do not lie on the forking causal path $\mathcal{F}$, and $\mathbf{p}^\star_{\overline{\mathcal{F}}}$ are the values they take on in the counterfactual local model $\mathcal{M}^{-V}(Ch)_{\mathbf{Dev}'}[Pa \rightarrow pa^*, \mathbf{P}_{\mathcal{F}} \rightarrow \mathbf{p}^*_{\mathcal{F}}]$, where $\mathbf{Dev}'$ is the subvector of $\mathbf{P}_{\mathcal{F}} \cup (Pa)$ which take on deviant values in $\mathbf{p}_{\mathcal{F}} \cup (pa)$ and default values in $\mathbf{p}^*_{\mathcal{F}} \cup (pa^*)$.) Either (A-1) $< Pa, Ch >$ is an active departure-return variable pair (relative to $\mathcal{F}$ and the designated contrasts); or (A-2) it isn't. If (A-1), then $(pa, pa^*)$ is deviant, rather than default, so $Pa \in \mathbf{Dev}'$, and $Pa$ does not affect the variables in $\mathbf{P}_{\overline{\mathcal{F}}}$ in the local model $\mathcal{M}^{-V}(Ch)_{\mathbf{Dev}'}$. So the parents of $Ch$ not on $\mathcal{F}$ will take on the same values in the local counterfactual model, even if $Pa$ is not set to $pa^*$,

$$\mathcal{M}^{-V}(Ch)_{\mathbf{Dev}'}[\mathbf{P}_{\mathcal{F}} \rightarrow \mathbf{p}^*_{\mathcal{F}}] \models \mathbf{P}_{\overline{\mathcal{F}}} = \mathbf{p}^\star_{\overline{\mathcal{F}}} \tag{13}$$

Let $v^*$ be the value of $V$ such that $v^* = \phi_V(pa^*)$. Then, since $Ch$ is $V$'s only causal child, it follows from (12) and (13) that in $\mathcal{M}$, $(ch, ch^*)$ will causally depend upon $(\mathbf{p}_{\mathcal{F}} \cup (v), \mathbf{p}^*_{\mathcal{F}} \cup (v^*))$. So the elongated $\mathcal{F} + V$[61] will be a forking causal path in $\mathcal{M}$. If (A-2), then the variables in $\mathbf{P}_{\overline{\mathcal{F}}}$ are not among $Pa$'s causal descendants, so (13) must hold. Again, let $v^*$ be the value of $V$ such that $v^* = \phi_V(pa^*)$. Then, since $Ch$ is $V$'s only causal child, it follows from (12) and (13) that in $\mathcal{M}$, $(ch, ch^*)$ will causally depend upon $(\mathbf{p}_{\mathcal{F}} \cup (v), \mathbf{p}^*_{\mathcal{F}} \cup (v^*))$. So the elongated $\mathcal{F} + V$ will be a forking causal path in $\mathcal{M}$. If (B),

---

[60] This is the directed forking path $\mathcal{F}$, minus the directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$, and plus the new directed edge $Pa \rightarrow Ch$.

[61] This is the directed forking path $\mathcal{F}$, minus the directed edge $Pa \rightarrow Ch$, and plus the new directed edges $Pa \rightarrow V$ and $V \rightarrow Ch$.

then $\mathcal{F}$ will also be a forking causal path in $\mathcal{M}$, since including the interpolated variable $V$ will not alter any of the causal dependence relationships amongst any of the variables other than $Pa$ and $Ch$. □

**Proposition 2.** Productive Causation$^\star$ *defines a model-invariant relation. That is: (a) given a causal model* $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$*, with* $U \in \mathcal{U}$*,* $\mathbf{C} \subseteq \mathcal{U} \cup \mathcal{V}$*,* $E \in \mathcal{V}$*, and* $U \notin \mathbf{C}$*,* $\mathbf{C} = \mathbf{c}$ *is a productive cause of* $E = e$ *in* $\mathcal{M}$ *iff* $\mathbf{C} = \mathbf{c}$ *is a productive cause of* $E = e$ *in* $\mathcal{M}^{-U}$*. And (b) given a causal model* $\mathcal{M} = (\mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E}, \mathcal{D})$*, with* $\mathbf{C} \subseteq \mathcal{U} \cup \mathcal{V}$*,* $E, V \in \mathcal{V}$*, and* $V \notin \mathbf{C} \cup (E)$*, if* $V$ *is inessential, then* $\mathbf{C} = \mathbf{c}$ *is a productive cause of* $E = e$ *in* $\mathcal{M}$ *iff* $\mathbf{C} = \mathbf{c}$ *is a productive cause of* $E = e$ *in* $\mathcal{M}^{-V}$*.*

*Proof.* For part (a): if $\mathbf{C} = \mathbf{c}$ is a productive cause of $E = e$ in $\mathcal{M}$, then, in $\mathcal{M}$, there is a forking causal path from $\mathbf{C}$ to $E$. The exogenous $U \in \mathcal{U}$ will not be on this forking path, so removing it will not affect any of the causal dependence relationships between any of the variables on the path. Nor will it affect whether any departure and return variables are active or deviant. So there will be a forking causal path in $\mathcal{M}^{-U}$. If $\mathbf{C} = \mathbf{c}$ was not a productive cause of $E = e$ in $\mathcal{M}$, then there was no forking causal path from $C$ to $E$. The exogenous $U \in \mathcal{U}$ is not on any directed forking path from $\mathbf{C}$ to $E$, and removing it will not affect any of the causal dependence relationships between any of the variables on any directed forking path from $\mathbf{C}$ to $E$, nor whether any departure and return variables are active or deviant. So removing $U$ will not create any new forking causal paths. So $\mathbf{C} = \mathbf{c}$ will not be a productive cause of $E = e$ in $\mathcal{M}^{-U}$. Part (b) follows immediately from **Lemma 2**. □

## References


Beebee, Helen. 2004. "Causing and Nothingness." In Collins et al. (2004), 291–308. [24]

Briggs, Rachael. 2012. "Interventionist Counterfactuals." *Philosophical Studies*, vol. 160: 139–166. [6]

Collins, John. 2004. "Preemptive Prevention." In Collins et al. (2004), chap. 4, 107–117. [24]

Collins, John, Ned Hall & L. A. Paul, editors. 2004. *Causation and Counterfactuals*. The MIT Press, Cambridge, ma. [51], [52], [53]

Dowe, Phil. 2000. *Physical Causation*. Cambridge University Press, Cambridge. [24], [29]

Fair, David. 1979. "Causation and the Flow of Energy." *Erkenntnis*, vol. 14: 219–50. [29]

Galles, David & Judea Pearl. 1998. "An axiomatic characterization of causal counterfactuals." *Foundations of Science*, vol. 3 (1): 151–182. [6]

Hall, Ned. 2004. "Two Concepts of Causation." In Collins et al. (2004), 225–276. [8], [21], [24], [29]

—. 2007. "Structural Equations and Causation." *Philosophical Studies*, vol. 132 (1): 109–136. [1], [8], [9]

Halpern, Joseph Y. 2008. "Defaults and Normality in Causal Structures." *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, 198–208. [1], [9], [13], [15]

—. 2016. *Actual Causality*. MIT Press, Cambridge, ma. [1], [9], [13], [15]

Halpern, Joseph Y. & Christopher Hitchcock. 2015. "Graded Causation and Defaults." *The British Journal for the Philosophy of Science*, vol. 66 (2): 413–457. [9]

Halpern, Joseph Y. & Judea Pearl. 2001. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." In *Proceedings of the Seventeeth Conference on Uncertainty in Artificial Intelligence*, John Breese & Daphne Koller, editors, 194–202. Morgan Kaufman, San Francisco. [1], [13], [14]

—. 2005. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." *The British Journal for the Philosophy of Science*, vol. 56: 843–887. [1], [13], [14], [38]

Hart, H.L.A. & Tony Honoré. 1985. *Causation in the Law*. Clarendon Press, Oxford, second edn. [33]

Hitchcock, Christopher. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese*, vol. 107 (3): 395–419. [18]

—. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy*, vol. 98 (6): 273–299. [1], [8], [13], [14], [38]

—. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review*, vol. 116 (4): 495–532. [1], [9], [13], [14]

Hitchcock, Christopher & Joshua Knobe. 2009. "Cause and Norm." *Journal of Philosophy*, vol. 106 (11): 587–612. [9], [34], [36]

Huber, Franz. 2013. "Structural Equations and Beyond." *The Review of Symbolic Logic*, vol. 6 (4): 709–732. [6]

Kahneman, Daniel & Dale T. Miller. 1986. "Norm Theory: Comparing Reality to Its Alternatives." *Psychological Review*, vol. 94 (2): 136–153. [9], [34]

Lewis, David K. 1973. "Causation." *The Journal of Philosophy*, vol. 70 (17): 556–567. [14], [16], [17], [18], [19]

—. 1986. "Causation." In *Philosophical Papers*, vol. II. Oxford University Press, New York. [3], [17], [37], [38]

—. 2004. "Causation as Influence." In Collins et al. (2004), chap. 3, 75–106. [8], [16]

Mackie, John L. 1965. "Causes and Conditions." *American Philosophical Quarterly*, vol. 2 (4): 245–55. [37], [38]

Maslen, Cei. 2004. "Causes, contrasts, and the nontransitivity of causation." In Collins et al. (2004), 341–357. [18]

Maudlin, Tim. 2004. "Causation, Counterfactuals, and the Third Factor." In Collins et al. (2004), 419–443. [9]

McDermott, Michael. 1995. "Redundant Causation." *The British Journal for the Philosophy of Science*, vol. 46 (4): 523–544. [18], [20], [24]

McGrath, Sarah. 2005. "Causation by Omission: A Dilemma." *Philosophical Studies*, vol. 123: 125–148. [9], [24], [36]

Menzies, Peter. 2004. "Causal Models, Token Causation, and Processes." *Philosophy of Science*, vol. 71 (5): 820–832. [1]

—. 2006. "A Structural Equations Account of Negative Causation." In *Contributed Papers of the Philosophy of Science Association 20th Biennial Meeting*. URL `http://philsci-archive.pitt.edu/2962`. [1]

Moore, Michael. 2009. *Causation and Responsibility*. Oxford University Press, Oxford. [33]

Paul, L. A. 2004. "Aspect Causation." In Collins et al. (2004). [18]

Paul, L. A. & Ned Hall. 2013. *Causation: A User's Guide*. Oxford University Press, Oxford. [9], [17], [21], [27]

Price, Huw. 1992. "Agency and Causal Asymmetry." *Mind*, vol. 101 (403): 501–520. [36]

Rosenberg, Ian & Clark Glymour. forthcoming. "Review of Joseph Halpern, Actual Causality." *The British Journal for the Philosophy of Science*. [43]

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton. [29]

—. 1994. "Causality without Counterfactuals." *Philosophy of Science*, vol. 61 (2): 297–312. [29]

Sartorio, Carolina. 2005. "Causes as Difference-Makers." *Philosophical Studies*, vol. 123: 71–98. [21]

—. 2007. "Causation and Responsibility." *Philosophy Compass*, vol. 2 (5): 749–765. [33]

—. 2016. *Causation & Free Will*. Oxford University Press, Oxford. [33]

Schaffer, Jonathan. 2000. "Causation by Disconnection." *Philosophy of Science*, vol. 67 (2): 285–300. [31]

Schaffer, Jonathan. 2005. "Contrastive Causation." *The Philosophical Review*, vol. 114 (3): 297–328. [18], [19]

—. 2012a. "Causal Contextualism." In *Contrastivism in Philosophy*, Blaauw, editor, chap. 2, 35–63. Routledge. [18]

—. 2012b. "Disconnection and Responsibility." *Legal Theory*, vol. 18 (4): 399–435. [33], [34]

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research*, vol. 66 (1): 81–103. [9]

Weslake, Brad. forthcoming. "A Partial Theory of Actual Causation." *The British Journal for the Philosophy of Science*. [1], [13], [15], [38]

Wolff, J. E. 2016. "Using Defaults to Understand Token Causation." *The Journal of Philosophy*, vol. 113 (1): 5–26. [9]

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford. [1], [13], [15], [38]

Yablo, Stephen. 2002. "De Facto Dependence." *The Journal of Philosophy*, vol. 99 (3): 130–148. [15]

—. 2004. "Advertisement for a Sketch of an Outline of a Prototheory of Causation." In Collins et al. (2004), chap. 5, 119–138. [15]