Structural Equations and Causation

Structural Equations and Causation

Ned Hall

§0 Introduction

Among philosophers and scientists interested in the nature of causation, one idea has gained a great deal of currency in recent years: that a proper understanding of the causal structure of any given situation can best be achieved by providing a *causal model* for that situation. Such a model will consist of appropriately chosen *variables*, together with *structural equations* that capture the relations of dependence among them. The key advantage to using these models—what, in the eyes of at least some authors, makes them indispensable—is that they provide tools by which to analyze, in a controlled and rigorous fashion, certain specialized counterfactuals in terms of which causation is to be defined.[1] Without the use of such models, so the story goes, a properly scientific understanding of causation will remain forever elusive.[2]

I think this is all a tissue of confusions. The sense among many philosophers of causation that the techniques of causal modeling constitute some exciting new advance is an overreaction to something whose legitimate pretensions are modest. At the same time, we can learn some fascinating lessons about causation by showing why.

Some of the reasons for a pessimistic response to structural equations approaches leap out once you focus on two obvious questions:

- What are variables?
- What are the truth-conditions for structural equations?

It's not so difficult to give sensible answers to these questions (which makes it all the more surprising that the literature doesn't contain any). But it's disappointing: what emerges is that far from being indispensable, causal models merely provide a useful means for selectively representing aspects of an *antecedently understood* counterfactual structure.[3]

A close look at the details of standard structural equations accounts of causation unearths further problems. Some, though worth pointing out, are of only local interest: the accounts suffer from obvious counterexamples; they fail to work as adver-

[1] For an example of a similar approach that—in my view, at least—*lacks* the needed controls, see Yablo (2004).

[2] For representative treatments of causation along these lines, see Pearl (2000), Hitchcock (2001), and Halpern & Pearl (2005).

[3] We'll also see that they are all too frequently used to *misrepresent* such structures; see in particular the discussion of 'late preemption' in §5.4, below.

tised when applied to some of the canonical preemption examples that have driven the causation literature. But a deeper problem remains, and it is quite interesting: typical accounts fail to incorporate a distinction between the *default* behavior of an object or system, and *deviations* therefrom.[4] (Very roughly: A system's default behavior is the behavior it would exhibit, if nothing acted on it. More helpful explanations will appear below!) This oversight is fatal; rectify it, and it becomes quite easy to produce a vastly improved structural equations account. (Perhaps better: a vastly improved account that could, if one liked, be presented within a structural equations framework.) So while much of the discussion to follow will be relentlessly critical, proceeding systematically through the problems just indicated, my hope is that this criticism will make vivid the central place of the default/deviant distinction in our thinking about causation. If so, then in debunking some of the hype surrounding the structural equations approach to causation, we will at least have found a pointer to an urgent and largely overlooked question: What makes the default/deviant distinction tick? I'll close with some tentative remarks about the larger significance of this topic.

§1 What is the aim of a structural equations account?

At various places in what follows, I'll be exhibiting specific cases, and trying to show that standard structural equations accounts deliver intuitively wrong verdicts about the causal structure of these cases. So it will pay to offer a brief preamble concerning the role of such intuitions. Now, I take it that a sensible attitude is *not* to treat firm intuitions about cases as absolutely non-negotiable "data"—*not*, that is, to adopt the kind of perspective one sees here, from Lewis:

> If one event is a redundant cause of another, then is it a cause *simpliciter*? Sometimes yes, it seems; sometimes no; and sometimes it is not clear one way or the other. When common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble. (Lewis 1986b, p. 194)

Why *not* accord intuitions about cases such a high degree of respect? Because a sensible metaphysical position is that facts about what causes what *reduce to* facts about the complete history of physical states the world occupies, together with facts about the fundamental laws that govern the evolution of these states. Deny this reductionist picture, and you might reasonably claim that close attention to causal intuitions will lead us to the deep, important metaphysical commitments embedded in

[4] I learned this useful terminology from Chris Hitchcock, whose own work on structural equations approaches clearly recognizes the importance of the default/deviant distinction. See also Maudlin (2004) for a very different approach that relies centrally on this distinction.

our ordinary causal thought and talk. But the reductionist picture already settles such commitments, at least as far as causation is concerned. Accept this reductionist picture—as I do, and as most authors in the counterfactual tradition seem to, either implicitly or explicitly[5]—and it seems that even perfect success at "triangulating" on intuitions about cases will accomplish nothing more than the production of a semantics for a fragment of English. Why should scientists, philosophers of science, or metaphysicians care about that?[6]

They shouldn't. But it doesn't follow that they should not care about intuitions *at all*. That would be an overreaction. Rather, they should treat intuitions about cases as defeasible evidence of the existence of a theoretically useful concept, worth careful articulation and study. This is, I think, a sensible attitude to take towards many topics in science and philosophy. Do our firmly held intuitive judgments involving the word "knowledge" track any concept of genuine interest for epistemology? Do our firmly held intuitive judgments involving the word "life" track any concept of genuine interest to biology? And so on. It's quite difficult to answer these questions well. But it seems clear that the best way to approach them is to *start out* with the assumption that trying to produce an account that respects the given intuitions will lead to something worthwhile. In the present case, we should hope that our intuitions about what causes what in a variety of specific cases lead to the production of a causal concept (or causal concepts—there needn't be just one!) that will do some interesting theoretical work. I will proceed in this hopeful manner, and make some proposals in §7 about the structure of one such causal concept.

The shift from viewing intuitions as non-negotiable data to viewing them as "guides" makes a difference to the dialectical role of examples. It won't do to exhibit some example, point out that the going structural equations approaches get it wrong, and declare them refuted. Rather, rejecting them on such a basis only makes sense if one can produce a *better* account, and say *why* it is better, beyond its ability to more closely fit the intuitive data. That is the standard of argument I will try to meet in what follows, by showing that there are systematic and interesting reasons why the structural equations accounts on offer do such a poor job (as they do) of reproducing our intuitive verdicts about cases.

[5] Even Lewis—the foregoing quote notwithstanding!

[6] Of course, this reductionist perspective about *causation* leaves wide open plenty of difficult and interesting deep-metaphysical questions, notably the "Humean" question whether facts about the fundamental laws themselves reduce to facts about the totality of physical states the world occupies. The point is just that no trace of *this* ontological dispute need carry forward into a dispute about the nature of causation. For more on this topic, see Hall (2004b), and Hitchcock (2003).

§2 Some simple examples

Before getting to the gory details, let's start with some examples that work very well to give the flavor of structural equations approaches, and on which they succeed rather admirably (although, as we'll see later on, for the wrong reasons!). We'll consider simple and undoubtedly familiar systems comprising interacting "neurons" (not the real thing, of course), that can fire if appropriately stimulated, and in firing send stimulatory or inhibitory signals to other neurons.

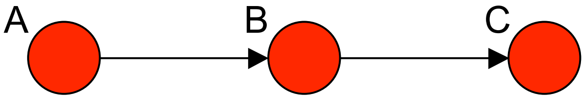


Figure 1

Here, for example, neuron A fires, sending a stimulatory signal to B, which fires as a result; B's firing sends a stimulatory signal to C, which fires as a result. The order of events, unless otherwise specified, is left-to-right. The firing of a neuron is indicated by shading its circle red, the presence of a stimulatory "channel" between two neurons by an arrow (the passage of a stimulatory signal is not explicitly represented). Throughout, I'll use capital letters interchangeably to refer to neurons and to events of their firing.

Of course, there is hardly any mystery about what causes what in a situation like that depicted in figure 1—nor in most other "neuron diagrams" (though some exceptions will appear later). And this is one of the advantages of working with such diagrams: they provide very clear test cases for any analysis of causation. But they also have an additional advantage, which is that they can help bring out what the differences between rival accounts of causation boil down to. That advantage is not on display in figure 1, because it's too simple. So consider instead figure 2:

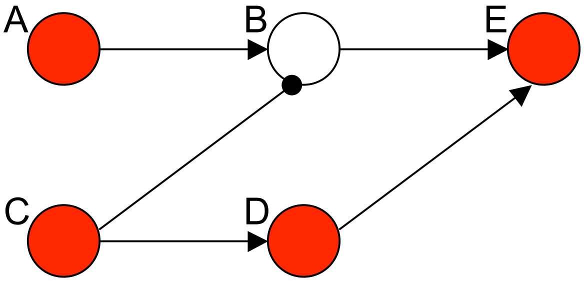


Figure 2

Here, A and C fire simultaneously. A sends a stimulatory signal to B; but at the same time, C sends an inhibitory signal to B. (The line with a blob on the end indicates an inhibitory channel.) Consequently, B does not fire—although it would have, had C not fired. E therefore fires, not as a result of any signal from B, but rather as a result of the signal from D, which fires as a result of the signal from C. The standard verdict about this case is that C is a cause of E, and A is not. Quite clearly, many real-world situations have this simple preemptive structure.

As is well-known, examples like figure 2 scotch the otherwise attractive idea that causation should be identified with *counterfactual dependence*: C is a cause of E iff had C not occurred, E would not have occurred. Since E in figure 2 does not thus depend on C, the account fails. But many have thought that the guiding idea behind it is correct, and we can usefully categorize various attempts to improve on the simple analysis by how they handle cases like figure 2. Here are the main options:

• Even though E does not depend on C, it *does* depend on D, and D on C[7]; combine the transitivity of causation with the claim that dependence at least *suffices* for causation, and you get the desired result that C is a cause of E. (See, most famously, Lewis 1973.)

• Even though E does not depend on C, it *does* "minimally" depend on a set containing C (namely, the set {A,C}), in the sense that had neither event in the set occurred, E would not have occurred, while the same is not true of any subset. Some (e.g. Ramachandran 1997) try to develop a counterfactual account that exploits this idea.

I won't consider these two approaches further (but see Hall & Paul 2003 for some criticisms). The next, however, will occupy us for much of the rest of the paper:

• E depends on C, *holding certain facts fixed*—in this case, the fact that B does not fire (at the relevant time). Yablo (2004) and Hitchcock (2001) take this approach, the latter within a structural equations framework. The approach of Halpern & Pearl (2005) yields this test as a special case.

Finally, towards the end of the paper I will outline a way to exploit the default/deviant distinction that may make the following approach viable:

• There is a process (viz., sequence of events) connecting C to E that has the right *intrinsic character* to qualify C as a cause of E (whereas no such process connect-

[7] With, of course, the usual understanding that the dependence is "non-backtracking": it's not that if D had not fired, that would have been because C did not fire, hence E would have fired all the same. Lewis (1979) gives what has come to be viewed as the standard treatment of non-backtracking conditionals. I think the influence this article has had is highly unfortunate, because its approach is badly confused. See §4.

ing A to E does); this can be brought out by examining the *counterfactual* structure of duplicates of this process, in suitable "test" circumstances. Hall (2004a) takes this approach, although as we'll see in §7, he now thinks that there may be a way to improve on that account.

Let's look at how a structural equations approach might handle figures 1 and 2. Now, we will shortly have to take up the questions about variables and structural equations raised above. But for the moment, we can ignore them, thanks to the highly sanitized nature of neuron diagrams. For example, in modeling figure 1 it is more or less obvious that we should choose three binary variables:

$\mathbf{A}$: has value 1 if neuron A fires (at the relevant time), 0 if it doesn't.

$\mathbf{B}$: has value 1 if neuron B fires (at the relevant time), 0 if it doesn't.

$\mathbf{C}$: has value 1 if neuron C fires (at the relevant time), 0 if it doesn't.

Note that there is nothing at all special about the numbers 0 and 1; they are mere labels.

Next, it is more or less obvious how to write down structural equations that capture the relations of immediate dependence between these variables:

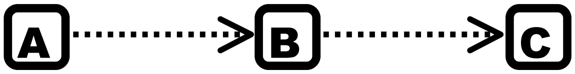
$$\mathbf{C} \Leftarrow \mathbf{B}$$

$$\mathbf{B} \Leftarrow \mathbf{A}$$

Thus, the first of these equations says, roughly, that C will fire iff B does. Note that I use "$\Leftarrow$" instead of the customary "=" because (as fans of structural equations regularly point out) the relation we mean to represent is *not* identity, but rather an asymmetric relation that captures the way in which the variable on the left-hand side has its value immediately *determined by* the values for the variables on the right-hand side (e.g., the variable $\mathbf{C}$ is to be "set" to the same value as $\mathbf{B}$). We'll look more closely at how to understand these equations in §4; for now, we can simply note that in general, for any variable $\mathbf{X}$ in any given model, the structural equations for that model will distinguish those other variables that $\mathbf{X}$ depends on (either immediately or mediately) from those it doesn't: $\mathbf{X}$ will depend on $\mathbf{Y}$ iff there is a sequence of variables $\mathbf{Y}$, $\mathbf{Z_1}$, $\mathbf{Z_2}$, …, $\mathbf{Z_n}$, $\mathbf{X}$ such that $\mathbf{Y}$ appears on the right-hand side of the structural equation for $\mathbf{Z_1}$, $\mathbf{Z_1}$ appears on the right-hand side of the structural equation for $\mathbf{Z_2}$, …, $\mathbf{Z_n}$ appears on the right-hand side of the structural equation for $\mathbf{X}$. There is thus a sharp distinction between *endogenous* variables, which depend on other variables, and *exogenous* variables, which don't. (E.g. in this model, $\mathbf{A}$ is the sole exogenous variable.)[8]

[8] A small technical nicety: equations must take the most "efficient" form—we can't, for example, make $\mathbf{B}$ here depend on $\mathbf{C}$ by rewriting the second equation as $\mathbf{B} \Leftarrow \mathbf{A} + \mathbf{C} - \mathbf{C}$. More exactly, we

We can give a partial but vivid representation of this system of equations/variables by means of the following *directed graph*:



Directed graph for figure 1

The graph tells us that **A** is an exogenous variable (relative to the given model), that the equation for **B** has **A** as its sole 'input', and that the equation for **C** has **B** as its sole input; hence this graph simply abstracts from the pair of structural equations given above. Despite its superficial similarity to figure 1, this directed graph should obviously not be *confused* with figure 1. (For example, only figure 1 contains a depiction of what *actually happens*.)

Virtually every structural equations account of causation will say the same thing about why A, in figure 1, is a cause of C, and will say it in terms of the proffered causal model. Here is the idea. In the situation as it actually unfolds, the variables take on these values:

A = B = C = 1

But the model allows us to consider what *would* have happened, had **A** had the value **0** (i.e., had A not occurred): we simply set

A = 0

and 'update' the values of **B** and **C** in accordance with the structural equations. We conclude:

if A = 0, then C = 0

It is because this conditional is true that A is counted a cause of C.

Fine, but why doesn't C likewise turn out to be a cause of A? Because of a further stipulation about how to evaluate these conditionals, one that doesn't kick in for the conditional just considered. Specifically, if we wish to evaluate

If X = v, then P

where **X** is some variable, **v** some possible value for it, and **P** some claim whose truth will be determined by the distribution of values for variables in whatever model we are using, then we must first distinguish those variables in the model that depend

can say that a variable **Y** in the equation for **X** is *irrelevant* iff, for each way of specifying the values of the *other* variables in the equation, there is a value **v** such that the equation guarantees that **X** = **v**, regardless of the value of **Y**. What we require is that no structural equation contain any irrelevant variables.

on $\mathbf{X}$ from those that don't. In evaluating the given conditional, the latter variables have their values *held fixed* at whatever they actually are; only the values of the former are updated in accordance with the structural equations. The total set of values that results then determines the truth of $\mathbf{P}$, and so the truth of the conditional. Since $\mathbf{A}$ does not depend on $\mathbf{C}$, we have

if C = 0, then A = 1

Hence C does not come out a cause of A.

Some comments.

First, in the more general treatment (given in Pearl & Halpern 2005, for example), the sort of thing that can be an *effect* is any proposition whose truth is determined by the distribution of values for variables in the model; the sorts of things that can be causes are arbitrary conjunctions of claims of the form "$\mathbf{X = v}$". I'll mostly ignore these extra complications, sticking to cases where what we wish to discern is the causal relationship, if any, between single events. Still, to get a story about such plain-vanilla event causation we clearly need some suitable translation of claims to the effect that some event occurs into, so to speak, the language of the model. Neuron diagrams are easy; but as we'll see later on, there are plenty of examples in which it is not so obvious how to do this translation.

Second, it's not actually guaranteed that a conditional—even if of the right form—will be *assigned* a truth-value by this recipe. Why not? Because, for all we've said, the system of structural equations for a given model might contain *loops*, so that distinct variables $\mathbf{X}$ and $\mathbf{Y}$ depend on *each other*. If so, it can happen that, for a particular choice $\mathbf{X = v}$, there is no way to update the values of the variables that depend on $\mathbf{X}$, consistent with the structural equations. This issue might matter, if we wished to use the structural equations approach to analyze situations involving backwards causation. We don't.[9] So I'll assume, henceforth, that our causal models behave themselves, and never feature such loops.

Third, this account of conditionals will of course remind you of the requirement, standard in counterfactual analyses of causation, that the counterfactuals used in the analysis be given a *non-backtracking* reading. You might therefore suspect the need for a story—perhaps involving Lewis's (1979) "miracles"—that will secure this reading. No such story is required. Once the structural equations are in place, the truth-conditions for these conditionals are perfectly well-defined. Now, it is a *further* question what the truth-conditions for these *structural equations* are, and it will probably

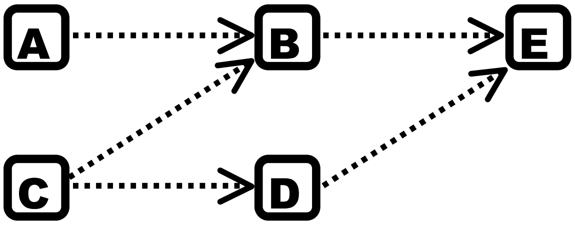
[9] The problem indicated here for accommodating backwards causation is not at all peculiar to the structural equations approach, but affects *any* counterfactual analysis. See Arntzenius & Maudlin (2005) for relevant details.

come as no surprise that answering that question will revive the issue of "miracles" (see §4).

Fourth, one might wonder whether the conditionals being analyzed *just are* ordinary English counterfactual conditionals. Pearl (2000) seems to think so, but the proposal doesn't really survive scrutiny. The main reason is that decent truth-conditions for the structural equations will, as we'll see, need to *rely on* counterfactuals; so as an analysis, the account would be circular. In addition, the proffered truth-conditions make explicit reference to *a specified model*, and nothing so far guarantees that a conditional that receives a truth-value relative to one model must receive *the same* truth-value relative to every other model that assigns it one. It would, of course, be rather embarrassing if this kind of stability of truth-values across models failed to obtain. We'll return to this issue in §4.

Finally, one might wonder what the big deal is, if the example of figure 1 is supposed to showcase the virtues of the structural equations approach. Isn't this just another counterfactual analysis of causation, with some pointlessly distracting talk of "models" and "equations"? Well, perhaps; but we can't deliver that verdict just yet. To be fair—and to see what the fuss is about—we need to look at the treatment of figure 2.

With the obvious choice of variables, here is the directed graph for the model we will use to analyze figure 2:



Directed graph for figure 2

And here are the structural equations:

$$\mathbf{E} \Leftarrow \mathbf{B + D - BD}$$
$$\mathbf{D} \Leftarrow \mathbf{C}$$
$$\mathbf{B} \Leftarrow \mathbf{A(1 - C)}$$

Finally, the actual values are these:

$$\mathbf{A = C = D = E = 1}$$
$$\mathbf{B = 0}$$

Here is one natural and attractive way to use this model to show that **C** is a cause of **E**, and **A** is not (adapted from Hitchcock 2001). First, observe that the sequence of variables **C-D-E** is, in an obvious sense, a *path* from **C** to **E**: i.e., a sequence such that each variable immediately depends on its predecessor in the sequence. Given this choice of path (not the only possible choice, obviously), **B** is an *off-path* variable. Next, even though the conditional

if C = 0, then E = 0

is false, the following conditional is *true*:

if (C = 0 & B = 0), then E = 0

It is because *this* conditional is true that C counts as a cause of E. Why is it true? Because of a natural generalization of the recipe given above: We look at the variables mentioned in the antecedent. We hold fixed the values of all variables that depend on neither of them. We update the values of the remaining variables by means of the structural equations. So **A**, which depends on neither **B** nor **C**, retains its value 1; the value of **D** is updated to **0** by the second equation; the value of **E** is updated to **0** by the first equation.

More generally, suppose we wish to determine whether event C is a cause of event E. We construct an appropriate causal model, with a (typically binary) variable **C** for C and **E** for E, and the customary values of 0 and 1. Then C is a cause of E just in case there is a path from **C** to **E**, such that for zero or more off-path variables $\mathbf{X}_1$, …, $\mathbf{X}_n$ with actual values $\mathbf{v}_1$, …, $\mathbf{v}_n$, the conditional

if (C = 0 & $\mathbf{X}_1$ = $\mathbf{v}_1$ & … & $\mathbf{X}_n$ = $\mathbf{v}_n$), then E = 0

is true. It's easy to check that this account not only delivers the verdict that C in figure 2 is a cause of E, but also the verdict that A is *not* a cause of E. Notice that if **E** depends on **C** *outright* (i.e., 'holding fixed' nothing), then C automatically qualifies as a cause of E.

Now, before proceeding we should pause to make a certain limitation explicit, which is that we have not offered an account of *causation* so much as an account of *causation-relative-to-a-model.* Halpern and Pearl are admirably forthright on this point, and worth quoting in detail:

> According to our definition, the truth of every claim must be evaluated relative to a particular model of the world; that is, our definition allows us to claim only that C causes E in a (particular context in a) particular structural model. It is possible to construct two closely related structural models such that C causes E in one and C does not cause E in the other. Among other things, the modeler must decide which variables (events) to reason about and which to leave in the background. We view this as a feature of our model, not a bug. It moves the questions of actual causality to the right arena—debating which of two (or more) models of the world is a better

> representation of those aspects of the world that one wishes to capture and reason about. (Halpern & Pearl 2005, p. REF)

It will emerge in §7 that there is *something* to be said for this perspective. But not much. And we'll also see, along the way, that Halpern and Pearl drastically overestimate its virtues. That's already somewhat obvious: for example, an account that fails to tell us that, in figure 2, C is a cause of E *simpliciter* does not deserve to be taken seriously. For now, we should take it that there is simply an important bit of unfinished business, which is to say what it is for a model to be *appropriate* for a given situation. Equipped with that distinction, we could say that C is a cause of E iff, for every appropriate model M of the situation in which C and E occur, C is a cause of E relative to M. Happily, for many of our examples it will be obvious what an appropriate model is.

The account just sketched displays but one of many options for using causal models to provide an account of causation. We can usefully contrast a second option, a simplification of the approach taken in Halpern & Pearl (2005). The first step is to liberalize the foregoing account, by allowing the off-path variables to take on *non-actual* values in the crucial conditional: C is a cause of E just in case there is a path from **C** to **E**, such that for zero or more off-path variables $\mathbf{X_1}$, …, $\mathbf{X_n}$ and (not necessarily actual) values $\mathbf{v_1}$, …, $\mathbf{v_n}$, the conditional

$$\textbf{if } (C = 0 \ \& \ X_1 = v_1 \ \& \ \ldots \ \& \ X_n = v_n), \textbf{ then } E = 0$$

is true. Of course, that's *too* liberal: for example, it counts A in figure 2 as a cause of E, and more generally counts preempted alternatives as genuine causes. So we add a further restrictive condition, which is that the following conditional must also be true:

$$\textbf{if } (C = 1 \ \& \ X_1 = v_1 \ \& \ \ldots \ \& \ X_n = v_n), \textbf{ then } P$$

where **P** 'says' that all of the variables on the chosen path from **C** to **E** have their *actual* values. The guiding idea is that C is a cause of E just in case there are some external contingencies that *could* have obtained, such that if they *had*, then (i) E would have depended on C; but (ii) the process connecting C to E would have been unaffected.[10]

Now A in figure 2 no longer gets counted a cause of E. There is but one path from **A** to **E**. The only off-path variable that matters is **C**, and the only value that matters is $\mathbf{C = 0}$. And while

$$\textbf{if } (A = 0 \ \& \ C = 0), \textbf{ then } E = 0$$

[10] Halpern and Pearl's extra condition is strictly weaker than (ii), allowing that the C-E process *could* have been altered by these external contingencies, so long as the alterations were in a specific sense irrelevant.

is true,

if (A = 1 & C = 0), then (A = 1 & B = 0 & E = 1)

is *false*.

Notice that this second account (henceforth: the "HP-account") is strictly more permissive than the first (henceforth: the "H-account"). It's an easy exercise to show that if the H-account calls C a cause of E, relative to a given model M, then the HP-account must also call C a cause of E, relative to M. To show the converse false, consider figure 3:

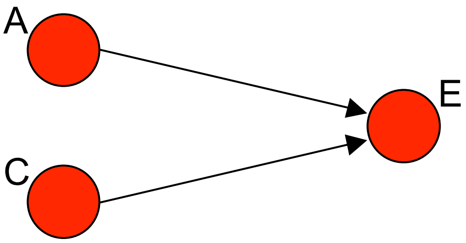


Figure 3

Here, the firings of A and C symmetrically overdetermine the firing of E. According to the H-account, neither A nor C is a cause of E; according to the HP-account, both are.[11] More examples on which the accounts differ will emerge in §5.

So far, the enthusiasm toward structural equations approaches might seem to be justified. At the very least, they appear to provide a novel, interesting, and effective means for treating certain of the preemption cases that have ever been the bane of counterfactual analyses of causation. Unfortunately, I think these appearances mislead. A proper assessment requires, to begin, a more careful, systematic statement of what causal models *are*. We'll take up that topic next.

§3 Variables

Neuron diagrams are easy to model, in large part because it is so easy to choose variables for them: For each neuron, and each relevant time-interval (roughly: each time interval such that that neuron could, in the circumstances, do something interesting in that time-interval), we introduce a variable to correspond to the state of that neuron during that interval. Typically, two values will suffice, one for "firing" and

[11] I take it that neither account scores a decisive victory over the other, on this example, since it's too unclear what a proper account of causation ought to say about symmetric overdetermination. One might even, in a pluralistic spirit, wish to accept *both* accounts, taking them to characterize slightly different but equally legitimate and useful causal concepts.

one for "not firing". But we can easily add values, if we wish to distinguish different *ways* the neurons can fire.

All of this suggests a more general prescription for choosing variables and values, for an arbitrary system we might wish to model: First, find a way to "carve up" the system into discrete, well-defined sub-systems. Second, for each relevant sub-system, and each relevant time or time-interval, introduce a variable to characterize the intrinsic physical state of that sub-system at that time, or during that time-interval.

I used "relevant" twice, partly because not every sub-system needs explicit representation (for example, one need not bother with variables corresponding to the stimulatory and inhibitory channels between neurons), and partly because not every moment of time is such that the behavior of the system at that time needs representing (for example, one also need not bother with variables that characterize the state of neurons before or after the events under consideration, or during the passage of signals).[12]

In addition, it may not always be straightforward how to "carve up" the given system into sub-systems. It will be *fairly* straightforward, if the system is constituted by a number of clearly distinguishable, interacting parts. But that won't always be the case—at least, at the desired level of description. Consider the flow of water down some rapids: what choice could we make of interacting parts, given that we don't wish to introduce variables for the state of each water molecule at each moment? Here a kind of default option suggests itself, which is that we choose variables to correspond to reasonably well-defined regions of space at different times, or regions of spacetime. The price of exercising this option is, in general, that no set of variables will stand out as uniquely appropriate.

Patently, what I've offered is far very from an exact recipe for determining the variables and values appropriate for modeling any given situation. But that is perhaps as it should be: within broad but non-trivial constraints, many choices are permissible. But some fans of structural equations approaches think that I have not been nearly permissive enough.[13] They hold that variables should be allowed to correspond to any family—any family *whatsoever*—of pairwise-incompatible propositions

[12] But how can we be sure that we aren't tacitly relying on our *antecedent understanding of the causal structure of the given situation* to decide what we should and shouldn't "bother" with? A legitimate worry, particularly if structural equations accounts aim to be *substantially* illuminating of the nature of causation—i.e., if they aim to be reductive (as they really ought, I think). But I think the worry can be gotten around: it will be enough to have a guarantee that, if C causes E relative to model M, then C causes E relative to any model M* that is richer than M in that it introduces additional variables for aspects of the given situation not represented in M. It's an embarrassment for both the H- and HP-accounts that they fail this extendability condition; see §§5.2 and 5.3.

[13] Halpern, personal communication.

(i.e., each distinct value of the given variable corresponds to a distinct member of the associated family). What I have in effect done is to restrict these families, requiring that each proposition in one of them be about the state of a particular physical system or region of space at or during a particular time or time-interval. Why not relax this restriction?

Emphatically *not* because doing so will lead to a view according to which it is facts, and not events, that are the causal relata. I find nothing at all attractive about the widespread view[14] that facts (by which let's just mean: true propositions) are too abstract or non-natural or inert or whatever to be causes or effects. The worry is more theoretically driven, and has two parts: First, without some constraints on variables like those I have suggested, structural equations approaches will fail for silly reasons. Second, no clear *advantages* accrue to relaxing the suggested restrictions. So we should stick with them.

The case for the second reason—that there is no pressing reason to allow variables to correspond to arbitrary sets of propositions—will play out over the rest of the paper. But the trouble with allowing such arbitrary correspondence is easy to see, and comes in two forms. Consider figure 4:

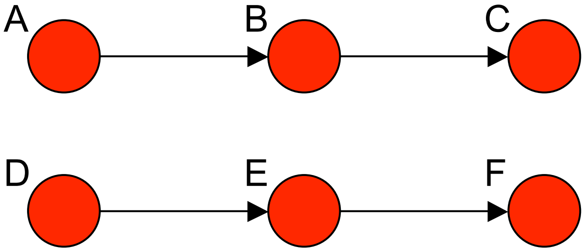


Figure 4

Here, two processes unfold concurrently. At time zero, A and D fire; a short while later B and E fire; then C and F fire. The two processes are completely causally isolated from each other. The natural causal model—the one that respected the constraints on variables I've suggested imposing—would of course reflect this fact. Here is a slightly *unnatural* causal model:

B: has value 1 if neuron B fires (at the relevant time), 0 if it doesn't.

C, **D**, **E**, **F**: defined similarly.

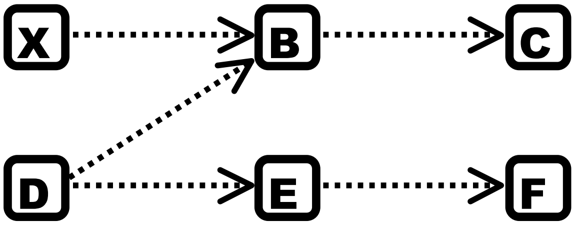
X: has value 1 if either both or neither of A and D fire; 0 otherwise.

[14] Not widespread among the causal modeling folk, to be sure; here I applaud their sensibility.

You can see where this is going, but we might as well let the gory details play themselves out. Accordingly, here are the equations:

$$\mathbf{C} \Leftarrow \mathbf{B}$$
$$\mathbf{F} \Leftarrow \mathbf{E}$$
$$\mathbf{E} \Leftarrow \mathbf{D}$$
$$\mathbf{B} \Leftarrow \mathbf{2XD - X - D + 1}$$

And here is the corresponding directed graph:



Directed graph for figure 4

All variables have value 1. Focus on the path **D-B-C**. Then

$$\textbf{if (D = 0 \& X = 1), then C = 0}$$

is true. So both the H-account and the HP-account classify D as a cause of C.

This result is plainly silly, and doesn't look any less silly if you insist that causal claims must always be relativized to a model. Figure 4 shows that this claim is just false. While there are, as we will see, *some* cases where, arguably, judgments about what causes what can be influenced by the decisions one makes about how to conceptualize the situation—decisions that can be reflected in a choice of model—this is simply not one of them: there is no sense whatsoever in which D is causally implicated in C. A much more honest response is to claim that the given choice of model is an inappropriate one, in some *objective* sense of "inappropriate". Fine, but what makes it inappropriate? Blindingly obvious answer: the choice of variable **X**, which, once it is introduced into the model, inevitably leads both structural equations accounts to conflate *logical* relations with *causal* relations.[15]

[15] Good grief; why is progress in philosophy so hard? This danger of conflating logical and causal relations is, after all, nothing new: it's been well-understood in the philosophy literature for at least 35 years, since Kim's 1971 critique of Mackie's account of causation.

Such choices can lead to a different kind of trouble as well, illustrated by a different bad causal model for figure 4. This time, we'll choose the sensible variables for A, C, D, and F, but collapse the behavior of B and E into a *single* variable:

$\mathbf{Y}$: has value 1 if both or neither of B and E fire, 0 otherwise.

Now the problem is not that we get the wrong verdicts about what causes what, but that we don't know how to set up the causal model in the first place. For, while one of the equations is easy:

$$\mathbf{Y} \Leftarrow \mathbf{2AD - A - D + 1}$$

the remaining equations—for $\mathbf{C}$ and for $\mathbf{F}$—are not. We might try the obvious

$$\mathbf{C} \Leftarrow \mathbf{A}$$

$$\mathbf{F} \Leftarrow \mathbf{D}$$

But these equations render $\mathbf{Y}$ entirely *irrelevant* to $\mathbf{C}$ and $\mathbf{F}$—which seems quite wrong, as $\mathbf{Y}$ concerns states of affairs not only temporally but causally *intermediate* between the events A and D, on the one hand, and the events C and F, on the other. What's more, the irrelevance of one variable for another *ought* to emerge as the result of applying some principled recipe for writing down structural equations, and then seeing that the *equation* renders the variable irrelevant.

So let's try to write down an equation for $\mathbf{C}$, say, in terms of $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{Y}$—bearing in mind that the resulting equation might reveal one or more of these variables to be irrelevant to $\mathbf{C}$ (in the mathematical sense: e.g., the equation will reveal $\mathbf{D}$ to be irrelevant iff for all values of $\mathbf{A}$ and $\mathbf{Y}$, the equation fixes a value for $\mathbf{C}$ independent of the value of $\mathbf{D}$). The equation can simply take the form of a table: there are eight ways to distribute values over $\mathbf{A}$, $\mathbf{D}$, and $\mathbf{Y}$, and for each way, we need to fix a value for $\mathbf{C}$. Some entries will be easy: for example, if $\mathbf{A = D = Y = 1}$, then we should have $\mathbf{C = 1}$. But other entries are not so easy. Suppose $\mathbf{A = D = 1}$, but $\mathbf{Y = 0}$; intuitively, this setting represents a situation in which, despite the fact that A and D both fire, one and only one of B and E fires. But which one? We will somehow have to settle that question, in order to figure out what the value of $\mathbf{C}$ should be. And there is manifestly no way to do so. So with the given choice of variables, we've destroyed our ability to construct a causal model.

Now, with respect to *this* problem the fan of permissiveness in variable choice could easily say, "Look, all this shows is that you have to choose variables in such a way as to make it possible to write down unambiguous structural equations. Duh!!" Indeed. And one of the advantages of *not* being so permissive about variable choice (over and above the need to avoid conflating logical and causal relations) is precisely that it makes the task of writing down unambiguous structural equations vastly easier. We'll see why in the next section.

§4 Structural equations and their truth-conditions

It's mildly scandalous that the causal modeling literature contains so little substantive discussion of what makes for a good choice of variables. But at least in practice, it's typically reasonably clear how to make this choice. By contrast, it is far, far more scandalous that the literature says so little about the truth-conditions for structural equations: scan Pearl's (2000) *book-length treatment* of causal modeling, for example, and you find nothing more substantive on this topic than scattered remarks such as the following: "The world consists of a huge number of autonomous and invariant linkages or mechanisms, each corresponding to a physical process that constrains the behavior of a relatively small group of variables." (p. 223) We are told nothing—*anywhere in the book*—that might yield an adequate understanding of such concepts as *autonomous*, *mechanisms*, or—most crucially—*constrains*.[16]

This lack of attention to such a crucial detail is shocking. At least we know what we *mean*—what sort of fact we are picking out—when we attribute some value to some variable; the issue with variables is the comparatively tractable one of what sorts of rules we should follow in collecting together a set of pair-wise incompatible propositions under the heading of a single variable. No such clarity attends the notion of a structural equation. A technical term—one of central importance in understanding what a causal model purports to represent—it nevertheless receives nothing approaching an adequate exposition. Lacking one, the notion cannot justly claim to provide the key to an illuminating understanding of causation.

Now, in suitably sanitized examples, a tacit and loose *counterfactual* interpretation of structural equations will naturally suggest itself, and may lead to a more or less obvious way to write down the structural equations. That is exactly what happened in the case of figures 1 and 2, for example. But we will see other examples—cases of preemption only slightly more complicated than that depicted in figure 2, for instance—in which such reliance on intuitive, non-explicit criteria for determining relations of dependency has led causal modelers badly astray. The need to avoid such mistakes provides us with a second reason to demand greater clarity.

I suspect that part of the tolerance for a lack of clarity about the truth-conditions for structural equations comes from the impression that there are no adequate alternatives to treating the notion as a kind of primitive, and using *it* to analyze the counterfactuals that will provide the ingredients for an account of causation. Now, to the

[16] See what happens when you send a scientist to do a philosopher's job?

extent that one thinks of Lewis's "miracles"-based account (in Lewis 1979) as providing the main or perhaps only alternative, this impression is entirely understandable. Without going into details, that account is baroque, poorly motivated, and fails even on its own terms.[17] But there is a much better account, which builds on a proposal in Maudlin (2003). It will work *quite* nicely to provide truth-conditions for structural equations—*provided* the strictures on choice of variables discussed in the last section are adhered to. To see how it works, and how it can be adapted to the needs of causal modeling, let's begin with a simple example.

At noon, Suzy throws a rock at a window. A few second later, the window breaks. Let C be her throw, E the breaking of the window. We seek truth-conditions for the conditional "if C had not occurred, then P", where P can be any claim about the post-C history of the world (e.g., "E does not occur"). So consider the state of the world—the complete, fundamental physical state of the world—at the time at which C occurs.[18] Consider a nomologically possible alternative to this physical state, which is just like it except that C does not occur. (Think of arriving at this state as follows: Begin with the actual state. Make localized changes to it—localized, that is, to the place or physical systems involved in C's occurrence—sufficient to guarantee C's non-occurrence.) The actual, fundamental physical laws proscribe a certain forward evolution for this physical state. If that forward evolution is such as to make it the case that P, then the conditional is true; if it is such as to make it the case that not-P, then the conditional is false. In the given example, we begin with an alternative state that is just like the actual state, save that Suzy doesn't throw. Forward evolution: the window doesn't break. In this way we secure the intuitively correct value *true* for "if Suzy hadn't thrown, the window would not have broken".

[17] A very quick example, which I learned years ago from Jim Woodward: Suppose that event E is multiply overdetermined by immediately preceding events D_1, …, D_n. Suppose that each of these events in turn is a joint effect of event C, which immediately precedes *them*. Such structures are transparently *possible*, if rare. On Lewis's account of non-backtracking counterfactuals, the "closest" world in which E does not occur is one that perfectly matches as much of actual history up to the time of E as possible, consistent with introducing a small, localized "miracle" (a violation of actual, though not counterfactual, law) that will throw history off course just enough to make E not occur. Now, one such miracle could be the non-occurrence of C. We could ensure a slightly *longer* stretch of perfect match of history, but only at the cost of n miracles—i.e., the non-occurrences of each of D_1 through D_n. Given the method that Lewis is unavoidably committed to for balancing the cost of miracles against the benefit of perfect match, we *must* choose the first option: a single earlier miracle, consisting in the non-occurrence of C. So, *in the non-backtracking sense that Lewis takes to suffice for causation*, it turns out that if E had not occurred, then C would not have occurred. That is a reductio. And there are others: see for example Elga (2001).

[18] *Begins* to occur, really—after all, C takes some *time* to occur. I'll omit this qualification henceforth.

We can also use this recipe to evaluate conditionals of the form “if C had occurred in such-and-such a manner, then P”: Begin with the actual state of the world at the time of C’s occurrence. Make localized changes, sufficient to make C occur in the specified way. Evolve the resulting state forward, in accordance with the actual fundamental laws. Check to see whether P comes true. Shortly, this version of the recipe will prove useful, in given systematic truth-conditions for structural equations.

A number of comments, before we proceed to that task.

First, these truth-conditions for counterfactuals don’t yet take proper account of indeterminism, of either the fundamental stochastic variety or the statistical mechanical variety.[19] We’ll ignore these issues here (fair enough, given that causal modelers routinely ignore them).

Second, it’s worth emphasizing that this account breaks sharply with philosophical tradition, in that it does not give a semantics for counterfactuals in terms of similarity between *possible worlds* but rather in terms of similarity between possible *complete physical states* of worlds. A smart move: much of what is so baroque about Lewis’s account, for example, flows from his insistence on defining a similarity relation between whole worlds. Note, in addition, that not much is required here of the notion of “similarity”: what we need, ultimately, is just a well-defined sense in which one complete physical state can be exactly the same as another, except in a certain specified respect that concerns a localized region or physical system.[20]

Third, we should not think that when we modify this localized region or physical system so as to make some actual event C fail to occur, we try to find an alternative state for this patch of the world that is as similar as possible to its actual state, consistent with the requirement that C not occur. That will lead to silly deliberations like the following: “Well, in the counterfactual situation in which Suzy’s throw does not occur, what happens instead? Does she perhaps toss it? But then how do we know that such a tossing is not numerically identical to the *actual* throw?” Lewis has some partially useful remarks on this point:

> What is the closest way to actuality for C not to occur? —It is for C to be replaced by a very similar event, one which is almost but not quite C, one that is just barely over the border that divides versions of C itself from its nearest alternatives. But if C is taken to be fairly fragile [i.e., characterized by stringent conditions of occurrence], then if C had not occurred and almost-C had occurred instead, very likely the effects of almost-C would have been much the same as the actual effects of C.

[19] Nor do they take account of the relativistic prohibition on talk of states-at-times. That oversight, I think, is easily fixed by switching to talk of states on appropriately chosen spacelike hypersurfaces.

[20] Quantum mechanical entanglement might raise a serious problem here. I’m going to blithely assume that the problem can somehow be solved. That’s okay, since if it *is* a problem I reckon it’s *everyone’s* problem.

> So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth value we thought it had. When asked to suppose counterfactually that C does not occur, we don't really look for the very closest possible world where C's conditions of occurrence are not quite satisfied. Rather, we imagine that C is *completely and cleanly excised from history*, leaving behind no fragment or approximation of itself. (Lewis 2004, p. REF; italics added)

I think Lewis's observations are right on target—up to the italicized portion, at which point they lapse into incoherence. What exactly does such "complete and clean excision" consist in? Removal of the event by some sort of metaphysical scalpel? Leaving behind … what? The Void?

A much better view is that for any given event, we work with an antecedently understood distinction between a *default state* for the region in which the event occurs, or for the physical system or systems to which it pertains. Conceiving of the event as one among various possible *deviations* from that default state, we answer the question, "What would have happened, had that event not occurred?" by returning the relevant region or system to its default state, holding the state of everything else fixed. It is in this way—and not by metaphysical surgery—that we fill in the Maudlin-recipe for evaluating counterfactuals. I'll say more about the default/deviant distinction in §6; for now, observe how well it fits with the way we naturally think about the case of Suzy and the rock, or indeed about neuron diagrams: In the case of Suzy, what we naturally think is that if she had not thrown the rock, what she would have been doing *instead* is standing there idly—*doing nothing*, as it were. Likewise, if we ask what would have happened, had a given neuron-firing not occurred, we naturally focus on an alternative situation in which that neuron remains, at the time in question, *in its dormant state*—not a situation in which it fires in a different manner, let alone a situation in which it is wholly absent, having been extracted in the course of surgery!

Fourth, it is not really to be hoped that—even with a default state specified—the recipe will yield a *unique* counterfactual state of the world. Here, multiple realization reigns, and we should correspondingly expect limits on what we can say about any given counterfactual situation. If Suzy's throw hadn't occurred, the window wouldn't have broken. –Not just *then*, at least. Would it have *remained* unbroken for the next year? We don't know, of course, and not because it's too hard to find out![21]

Fifth, the recipe is quite limited in scope. It says nothing about conditionals such as the following: "If gravity had obeyed an inverse-cube law, Kepler's second law still would have held." (True, by the way.) Nor is it built to handle "backwards" condi-

[21] There's room for fussing here about the relation between "might" and "would" counterfactuals. We can all agree that if Suzy had not thrown, the window might have remained unbroken for the next year. Lewis (1973b) argues that we must therefore *deny* that the window would have broken in that time, and not merely claim not to know whether it would. Nothing hangs on this issue, here.

tionals, in which the consequent concerns a time or times *before* the time or times that the antecedent is about.[22] Neither limitation poses a problem, given the purposes to which we will put the recipe; the second, in fact, is exactly what allows us to avoid talk of "miracles".[23] A further limitation is even more clearly beneficial: Referring back to figure 4, the recipe equivocates if we ask, for example, what would have happened if exactly *one* of A and D had fired; for we don't have a rule for deciding which one, and so can't specify in enough detail the relevant counterfactual state of the world. For some purposes, this won't matter: e.g., we can at least say that exactly one of C and F would have fired. For other purposes it will: e.g., we can't say that C would have fired, or that it would not have fired. Now, I think of this last limitation as a feature, not a bug: it will force us to be scrupulous in our choice of variables, and so gives content to the idea floated in the last section, that this choice needs to be made in a way that will allow for clean, unambiguous structural equations linking these variables.

Time to consider how to arrive at such structural equations. We'll start with a simple idea, spot the need for an amendment, and refine accordingly.

As an illustration, suppose we have some situation for which we wish to provide a causal model, and suppose we've decided that this model should make use of five variables: $\mathbf{C_1}$, $\mathbf{C_2}$, $\mathbf{D_1}$, $\mathbf{D_2}$, and $\mathbf{E}$. Respecting the strictures laid out in §3, we have chosen these variables in such a way that each has a well-defined time or time-interval associated with it. Let's suppose that $\mathbf{C_1}$ and $\mathbf{C_2}$ concern the same time, as do $\mathbf{D_1}$ and $\mathbf{D_2}$; let's suppose further that the temporal order among the five variables is this: $\mathbf{C_1}, \mathbf{C_2} < \mathbf{D_1}, \mathbf{D_2} < \mathbf{E}$. Then a simple approach is to stipulate that the equations for $\mathbf{D_1}$ and $\mathbf{D_2}$ shall include only $\mathbf{C_1}$ and $\mathbf{C_2}$, and that the equation for $\mathbf{E}$ shall include only $\mathbf{D_1}$ and $\mathbf{D_2}$. The Maudlin-recipe applies straightforwardly. To fix an equation for $\mathbf{D_1}$, for example, we need to determine, for each setting of the $\mathbf{C}$-variables $\mathbf{C_1 = v_1}$ and $\mathbf{C_2 = v_2}$, a resulting value for $\mathbf{D_1}$. Begin with the state of the

[22] Why not? Why not just take the alternative state, and evolve it *backwards* (in accordance with the actual fundamental laws), as well as forwards? Won't that give us truths about how the past would have been, had some present event not occurred, or occurred differently? In principle, yes. But in practice it will, for typical cases, be virtually impossible to say in any detail how the past would have been—even if the fundamental laws are time-reversible. (Try some examples, if you're not convinced.) Add the fact that our purposes here don't require us to extend the Maudlin-recipe in this way, and we should conclude that there is no real point to doing so.

[23] What about backwards causation? It can be accommodated, if one wishes, by some refinements that I won't go into here. Observe, though, that Lewis's "miracles" account provides—his claims to the contrary notwithstanding—no help in securing the possibility of backwards causation. See fn. 9.

world at the time the $\mathbf{C}$-variables concern. Modify it locally so as to make $\mathbf{C_1 = v_1}$ and $\mathbf{C_2 = v_2}$. Evolve the resulting state forward in accordance with the actual fundamental laws. The value for $\mathbf{D_1}$ will be that unique value $\mathbf{w}$ such that the proposition $\mathbf{D_1 = w}$ is guaranteed to be true, given this forward evolution.[24]

(What if there is no such unique value? Well, there won't be *more* than one; the worry is that there might not be *any*. If so, that shows that there was something wrong with our choice of variables—e.g., they weren't "fine-grained" enough. (We'll see an example in the next section, during the discussion of late preemption.) Then we should simply fix the problem, and move on.)

This account of the truth-conditions for structural equations almost works. But there is a problem, one that arises if we are, as it were, too parsimonious in our choice of variables. Consider figure 2 again. Suppose we simply *omit* the variables $\mathbf{A}$ and $\mathbf{B}$, choosing to construct a model using only $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$. Then the account just given will yield these structural equations:

$$\mathbf{E} \Leftarrow \mathbf{D}$$
$$\mathbf{D} \Leftarrow \mathbf{C}$$

There's no problem with the second, nor—in a *certain* sense—with the first; that is, the first correctly captures the way that $\mathbf{E}$ immediately depends on $\mathbf{D}$. But put them together in a single causal model, and that model will tell us (in accordance with the recipe presented in §2) that the conditional

$$\textbf{if C = 0, then E = 0}$$

is *true*. Now, no one said that *that* conditional—which, remember, is essentially a technical device explicitly defined by reference to a causal model—had to have, relative to any model that assigns it a truth-value, the *same* truth-value as the conditional "if $\mathbf{C}$ = 0, then $\mathbf{E}$ = 0" that we evaluate using the Maudlin-recipe. The latter conditional is straightforwardly *false*—false *full stop*, and not merely relative to this or that model. Perhaps we should rest content with the position that the former conditional can be true relative to some models (e.g., this one), and false relative to others (e.g., the more complete model of figure 2 provided in §2).

But that would be a mistake, a move that would shift even more of the burden of providing an adequate structural equations account of causation onto the project of producing the as-yet unwritten rules for choosing "appropriate" causal models. It's

[24] Don't be fooled, of course, by the heuristic talk of "modifying" and "evolving" (as if complete physical states were something we could manipulate). When cleansed of such talk, it's apparent that what we have provided here is a purely *metaphysical* story about what makes a given structural equation correct.

much better to lay down rules that guarantee a certain kind of *stability* in our causal models, so that a conditional like the foregoing one will receive the same truth-value relative to every model that assigns it one. There is a natural way to achieve this effect, one with the added benefit of guaranteeing that this truth-value will *match* the one yielded by the Maudlin-recipe.

Return to figure 2, and our overly parsimonious model for it that used only variables $\mathbf{C}$, $\mathbf{D}$, and $\mathbf{E}$. The trouble we got into with this model derived from the fact that, in the counterfactual situation in which $\mathbf{C}$ has the value 0, one *consequence* of its having this value is that neuron B *fires*, which in turn guarantees that E fires. But our paired-down model contains no variable whose value could reflect the fact that B fires. One bad solution to this problem is to insist that an acceptable model contain a comprehensive enough set of variables, so that any relevant consequence of one variable's having a given value gets explicit representation in the values of other variables. I think that places too high a demand on the causal modeler, and at any rate there is a cleaner approach. To illustrate, I'll stick to this three-variable model for figure 2.

The temporal order of the variables is $\mathbf{C} < \mathbf{D} < \mathbf{E}$. In writing down equations for these variables, we adopt the policy that for a given variable, *any* temporally prior variable is allowed to figure in its structural equation. Since $\mathbf{C}$ is the sole variable prior to $\mathbf{D}$, we recover, using the Maudlin-recipe, the same equation for $\mathbf{D}$ as before:

$$\mathbf{D} \Leftarrow \mathbf{C}$$

But $\mathbf{E}$ is now allowed to functionally depend on both $\mathbf{C}$ and $\mathbf{D}$. That means that, for each of the four ways of assigning values to $\mathbf{C}$ and $\mathbf{D}$, we need to determine a resulting value for $\mathbf{E}$. So consider the case $\mathbf{C = x}$ and $\mathbf{D = y}$. Focus on the state of the world at the time that $\mathbf{C}$ concerns. Make local changes, sufficient to guarantee that $\mathbf{C = x}$. (If $\mathbf{C = x}$ in actuality, no changes will be necessary.) Evolve the resulting state forward *until the time that* $\mathbf{D}$ *concerns*. Make local changes to *this* state, sufficient to guarantee that $\mathbf{D = y}$. (Again, no changes may be necessary.) Evolve this newly modified state forward in time. Some value for $\mathbf{E}$ will result. That is the value that the structural equation for $\mathbf{E}$ should specify as output, when given as input the values $\mathbf{C = x}$, $\mathbf{D = y}$.

Let's test this approach. For the situation depicted in figure 2, the actual values $\mathbf{C = 1}$, $\mathbf{D = 1}$ obviously map to $\mathbf{E = 1}$. Given the values $\mathbf{C = 1}$, $\mathbf{D = 0}$, we begin with the (actual) state in which both C and A fire, evolve forward into a state in which B doesn't fire but D does, *locally modify* this state so that D *does not* fire (and B still doesn't), evolve *this* state forward, and see that E does not fire. So, for $\mathbf{C = 1}$, $\mathbf{D = 0}$, we must have $\mathbf{E = 0}$. It's routine to check the other two cases: $\mathbf{C = 0}$, $\mathbf{D = 1}$ gives us $\mathbf{E = 1}$; $\mathbf{C = 0}$, $\mathbf{D = 0}$ gives us $\mathbf{E = 1}$. More simply:

$$\mathbf{E} \Leftarrow \mathbf{1 - C + CD}$$

Notice, finally, that the fact that this is the correct equation for $\mathbf{E}$ depends crucially on what variables are included in the model. Reintroduce $\mathbf{B}$, for example, and the correct equation renders $\mathbf{C}$ irrelevant. That result, of course, is exactly as it should be.

The generalization of this recipe is straightforward: Suppose we have some variable $\mathbf{X}$ in some model M. If we have been scrupulous in our choice of variables, there will be a clear-cut distinction between those other variables in M that are temporally prior to $\mathbf{X}$, and those that are not. For each way of assigning values to the former variables, we can follow the 'sequential updating' variant of the Maudlin-recipe to fix a resulting value for $\mathbf{X}$. In this way, the fundamental laws, together with the actual history of the world, will fix a unique structural equation for $\mathbf{X}$ (in terms of the other variables in M).

Now consider some variable $\mathbf{C}$ in M that is temporally prior to $\mathbf{X}$. Consider the counterfactual situation in which $\mathbf{C} = \mathbf{c}$, arrived at by locally modifying the state of the world at the time that $\mathbf{C}$ concerns so as to make $\mathbf{C} = \mathbf{c}$, and evolving this state forward in time. This forward evolution will yield some assignment of values to all variables in the model: those that are temporally prior or concurrent with $\mathbf{C}$ will receive their *actual* values; the remaining variables may receive different values. What's more, given our truth-conditions for structural equations, the value that $\mathbf{X}$ receives in this counterfactual situation *must be the same* as the value that the structural equation for $\mathbf{X}$ yields, when given as input the values that all the variables prior to $\mathbf{X}$ receive, in this counterfactual situation. It follows that the conditional

if C = c, then X = x,

evaluated by the procedure described in §2, must, regardless of the details of the model M, receive the same truth-value as the counterfactual "if $\mathbf{C} = \mathbf{c}$, then $\mathbf{X} = \mathbf{x}$" (when evaluated by the Maudlin-recipe). We have thus arrived at truth-conditions for structural equations that are not only clear, but that also guarantee that the apparently *model-relative* truth-conditions for conditionals laid out in §2 are in fact *not* model relative: Any model that assigns a conditional a truth-value will assign it the same truth-value, and moreover will assign it the truth-value it *ought* to have (i.e., the truth-value determined by the Maudlin-recipe).

(We haven't covered the case of the more complicated conditionals used in both the H-account and the HP-account to define causation. No matter: it's an easy exercise to show that parallel results apply to them. What's more, we'll shortly see that it was a mistake to rely on such conditionals, in the first place.)

One final issue needs to be addressed, that concerns a certain problem that can arise from a combination of incautious choice of variables together with lake of care in writing down structural equations. Here is the abstract form of the problem: Suppose we have an equation for **B** in terms of **A**, and an equation for **C** in terms of **B**. Given our first equation, the following conditional is true:

if A = x, then B = y

Given our second, the following conditional is true:

if B = y, then C = z

Assuming the two equations in our three-variable model are correct, we can immediately conclude that

if A = x, then C = z[25]

But this reasoning will be fallacious if the meaning of "**B = y**", as it appears in the consequent of the first conditional, is not the same as the meaning of "**B = y**", as it appears in the antecedent of the second. Granted; but you might wonder what the worry is. After all, how could anyone be so stupid as to equivocate in this way? You'd be surprised.

The danger of equivocation becomes quite real when we introduce a binary variable X to represent some event, and stipulate that "X = 0" means that the event does not occur. Such a claim can be quite open-ended, in a way that will cause the modeler headaches, if she's not sufficiently careful. An example will bring out the potential problems nicely:

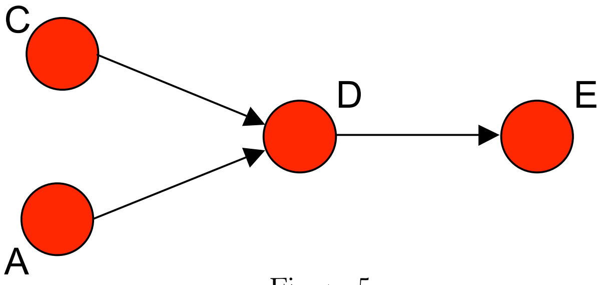


Figure 5

[25] It doesn't follow—and *better* not!—that transitivity for these conditionals holds in general. What secures transitivity in this case is that, given the facts of the case, the correct equation for **C** renders **A** irrelevant.

The connection between A and D in figure 5 is not of the normal kind; in particular, whether A fires *never* has an effect on whether D fires (even if C does not fire). No, D will fire iff stimulated by C. What A does is to determine whether D fires with normal intensity, as in figure 5, or feebly, as in figure 6:

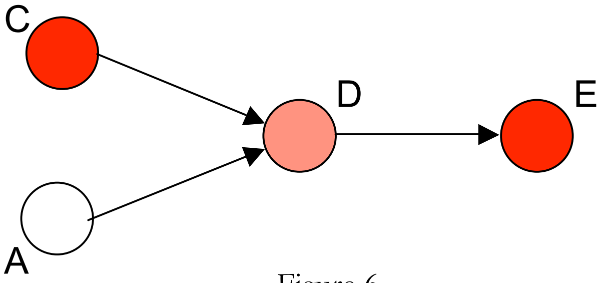


Figure 6

Neuron E, finally, will fire iff stimulated by D; what's more, the way in which it fires is completely insensitive to whether D fires feebly.

A sensible causal model for figures 5 and 6 will, obviously, introduce binary variables for A, C, and E, and a *ternary* variable for D. But suppose we don't do that; instead, we introduce a binary variable for D, stipulating that $\mathbf{D} = \mathbf{1}$ iff D fires normally, and $\mathbf{D} = \mathbf{0}$ iff it does not fire normally—i.e., either fires feebly or not at all. Then we can certainly write down this correct equation for $\mathbf{D}$:

$$\mathbf{D} \Leftarrow \mathbf{AC}$$

But we can't write down *any* correct equation for $\mathbf{E}$. The reason is that, in following the sequential updating version of the Maudlin-recipe, there will be cases in which we need to make local adjustments to the state of the world, at the relevant time, in order to guarantee that $\mathbf{D} = \mathbf{0}$; and we simply won't know *how* to make them. (Specifically, suppose we want to know what the equation for $\mathbf{E}$ should yield as value, for the inputs $\mathbf{A} = \mathbf{1}$, $\mathbf{C} = \mathbf{1}$, $\mathbf{D} = \mathbf{0}$.)

On the other hand, we could stipulate that $\mathbf{D} = \mathbf{0}$ iff D does not fire *at all* (at the relevant time), and $\mathbf{D} = \mathbf{1}$ iff it fires in some way or other. Then we can write down correct equations for both $\mathbf{D}$ and $\mathbf{E}$:

$$\mathbf{E} \Leftarrow \mathbf{D}$$

$$\mathbf{D} \Leftarrow \mathbf{C}$$

Now the problem is that we've lost the ability to model A's effect on D.

Finally, we could stipulate that $\mathbf{D} = \mathbf{0}$ iff D does not fire at all, and $\mathbf{D} = \mathbf{1}$ iff it fires feebly (alternatively: iff it fires normally). Then the foregoing equation for $\mathbf{E}$ is correct, but no equation for $\mathbf{D}$ is possible. Inevitably, then, something is left out: either we can't properly get E into the picture, or we can't properly get A into the picture, or we can't properly get D into the picture. The fix, obviously, is to let **D** have *three* distinct values: 0 (doesn't fire), 1 (fires feebly), 2 (fires normally). Then we get perfectly adequate equations:

$$\mathbf{E} \Leftarrow \mathbf{(3D - D^2)/2}$$

$$\mathbf{D} \Leftarrow \mathbf{C(A + C)}$$

The final point to emphasize is this: It would be a *total disaster* if we built a model that gave a role to both A and E, *not* by letting **D** have three values, but rather by stringing together the equation for **D** taken from the first model with the equation for **E** taken from the second:

$$\mathbf{D} \Leftarrow \mathbf{AC}$$

$$\mathbf{E} \Leftarrow \mathbf{D}$$

Now we really have equivocated, for the first equation is only correct if "**D = 1**" means that D fires normally, whereas the second is only correct of "**D = 1**" means that D fires somehow or other. Alas, we'll see just this mistake being committed, in the standard structural equations treatment of late preemption (§5.4).

We now have reasonably good answers to the two questions raised at the outset: What are variables? What are the truth-conditions for structural equations? Certainly, our answers are good *enough* to let us go ahead and use causal models in giving an account of causation, without feeling too ashamed of ourselves. Still, it should also be clear that *no special advantage accrues to using causal models*. We now understand what structural equations say—and understanding it, we can see that they are *nothing more* than a device for selectively representing aspects of an *antecedently understood* counterfactual structure. In any given situation, that structure is fixed by what happens, together with the fundamental laws, in a way that is articulated in a perfectly detailed and adequate manner by the Maudlin-recipe (including the sequential updating variant). And that fact gives us excellent grounds for concluding that, when it comes to producing a rich and illuminating account of causation, anything that can be accomplished by means of causal models can be accomplished just as straightforwardly without them. We'll see in §7 that this conclusion is exactly right.

§5 Trouble cases

Now we'll see how little is in fact accomplished by the structural equations accounts that have been put forward. (At least, the two sketched in §2; there won't be any profit in exploring the variants that have appeared in the literature.) We'll look at four more examples.

§5.1 Switches

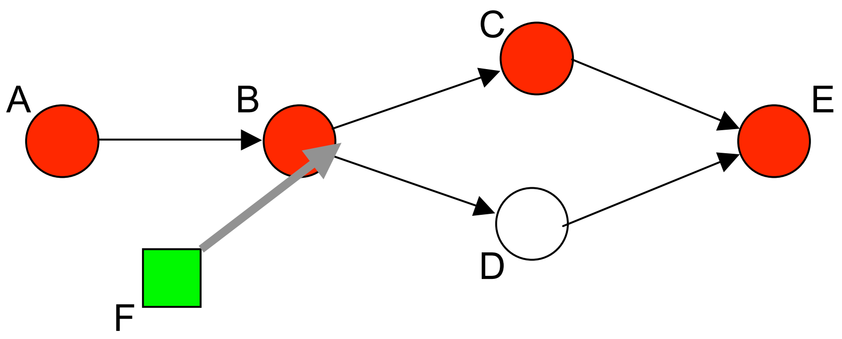


Figure 7

All of the neurons depicted here are normal, except for F. It's firing has no effect on the firing of B; rather, what F does is to determine down which of the two channels exiting from B the stimulatory signal from B travels. If F fires, as it does in figure 7, then the stimulatory signal gets sent to C; if it doesn't, the signal gets sent to D:

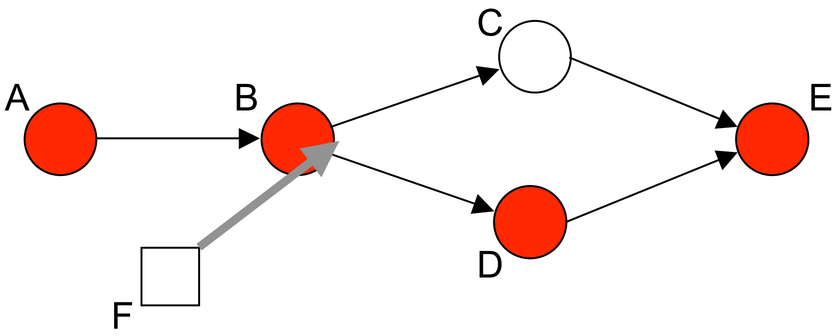


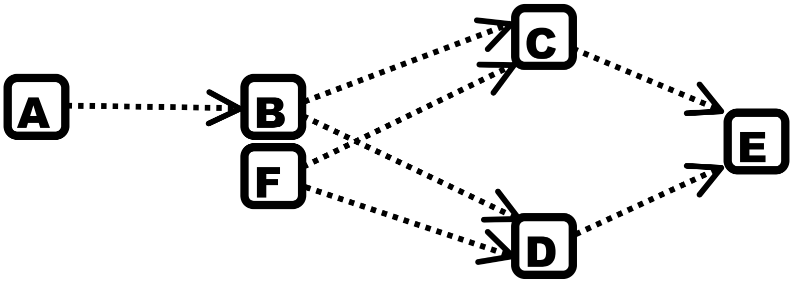
Figure 8

Neuron F thus acts as a "switch". Real-world analogues are easy to come by. For example, the following case, and variants, have been much discussed:

The Engineer: An engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the left-hand track, instead of the right. Since the tracks reconverge up ahead, the train arrives at its destination all the same.

Many people, myself included, share the judgment that such "switching" events are not causes of the relevant effects: F, in figure 7, is not a cause of E—

notwithstanding that it *is* a cause of C, and C of E.[26] Both the H-account and the HP-account say otherwise; it will be enough to look at the H-account to see why. Let's begin with the obvious causal model, which has this directed graph:



Directed graph for figures 7 & 8

Here are the equations (I'll leave it as an exercise to check that they are the correct equations, given the results of the last section):

$$\mathbf{E} \Leftarrow \mathbf{C} + \mathbf{D} - \mathbf{CD}$$
$$\mathbf{D} \Leftarrow \mathbf{B(1 - F)}$$
$$\mathbf{C} \Leftarrow \mathbf{BF}$$
$$\mathbf{B} \Leftarrow \mathbf{A}$$

Finally, the variables have these values:

$$\mathbf{A = B = C = F = E = 1}$$
$$\mathbf{D = 0}$$

One path from **F** to **E** is **F-C-E**. Then **D** is an off-path variable. Furthermore,

$$\textbf{if (F = 0 \& D = 0), then E = 0}$$

is true. So F turns out to be a cause of E. The same result holds in The Engineer: for the right-hand track is *in fact* empty; and if, in the counterfactual situation in which she doesn't flip the switch, it somehow remains so, then the train does not arrive at its destination.

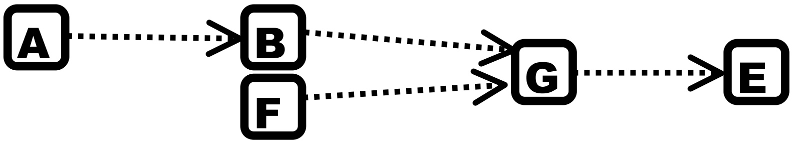
I do not think this result is *completely* devastating. (After all, I'm on record as providing not-very-compelling but not-entirely-worthless reasons for thinking that F *is* a cause of E.) But it *is* a problem. And, there is—as we will shortly see—a natural way to develop an account that avoids it. At any rate, there is no clean way around it, on the approaches we're currently considering. Now, Halpern and Pearl think otherwise.

[26] In Hall 2000, I labored mightily to have the contrary intuition, in order to preserve the transitivity of causation. I now think that was probably a mistake.

They point out, in effect, that we *could* have described figures 7 and 8 by means of a model that collapses the distinct variables **C** and **D** into one:

G: has value 2 if neuron C fires and D does not; has value 1 if D fires and C does not; has value 0 if neither D nor C fire.

This choice yields a modified directed graph:



Modified directed graph for figures 7 & 8

And new structural equations:

$$E \Leftarrow (3G - G^2)/2$$

$$G \Leftarrow B(B + F)$$

$$B \Leftarrow A$$

It's routine to verify that, relative to *this* model, both the H-account and the HP-account deliver the verdict that while A and B are causes of E, F is not.

Halpern and Pearl commend these contrasting verdicts. Commenting on The Engineer, they write:

> Can a change in representation turn a non-cause into a cause?
>
> It can and it should! The change to a two-variable model [i.e., the model that distinguishes **C** from **D**] is not merely syntactic, but represents a profound change in the story. The two-variable model depicts the tracks as two independent mechanisms, thus allowing one track to be set (by action or mishap) to false (or true) without affecting the other. Specifically, this permits the disastrous mishap of flipping the switch while the left track is malfunctioning. More formally, it allow a setting where $F = 1$ and $C = 0$. Such abnormal settings are imaginable and expressible in the two-variable model, but not in the one-variable model. Of course, if we disallow settings where $F = 1$ and $C = 0$, or where $F = 0$ and $D = 0$, then we are essentially back at the earlier [i.e., one-variable] model. The potential for such settings is precisely what renders F a cause of E in the model of figure 7.
>
> Is flipping the switch a legitimate cause of the train's arrival? Not in ideal situations, where all mechanisms work as specified. But this is not what causality (and causal modeling) are all about. Causal models earn their value in abnormal circumstances, created by structural contingencies, such as the possibility of a malfunctioning track. It is this possibility that should enter our mind whenever we decide to designate each track as a separate mechanism (i.e., equation) in the model and, keeping this contingency in mind, it should not be too odd to name the switch position a

> cause of the train arrival (or non-arrival). (Halpern & Pearl 2005, p. REF, with minor relettering)

I think the discussion here has gone off the rails. (Sorry!) Let's first focus on this sentence, with italics added: "The two-variable model depicts the tracks as two independent mechanisms, thus *allowing* one track to be set (by action or mishap) to false (or true) without affecting the other." The ill-chosen word "allowing" suggests that they think that it is *only* by introducing distinct variables for the two tracks that one can, for example, represent "the disastrous mishap of flipping the switch while the left track is malfunctioning". This suggestion is, of course, absurd. The one-variable model (i.e., with **G** in place of **C** and **D**) will serve just fine, provided we give **G** more values. For example, nine of them:

G: has value 0 if the train is sent down neither track; 1 if sent down left track and both tracks operational; 2 if sent down right track and both tracks operational; 3 if sent down left track and left track alone malfunctions; 4 if sent down right track and left track alone malfunctions; 5 if sent down left track and right track alone malfunctions; 6 if sent down right track and right track alone malfunctions; 7 if sent down left track and both tracks malfunction; 8 if sent down right track and both tracks malfunction.

Notice, in addition, that the 'enhanced' one-variable model that incorporates **G** does *not* deliver the verdict that the switch-flipping is a cause of the arrival—despite the addition of six new values for **G**. So it was not the choice to recognize certain 'structural contingencies' that altered the verdict about the causal status of the switch-flipping; rather, it was the choice to represent these contingencies *in a certain way*—a way, moreover, that is *purely optional*. Presumably, *something* ought to govern this further choice, something, that is, beyond the desire to model the given contingencies. But what that something is is left a complete mystery.

What's more, there are, quite plainly, *other* reasons for choosing to introduce distinct variables besides the desire to represent outlandish contingencies. How about this simple one: the distinct variables correspond to distinct physical parts of the total system in question? In §3, we saw the pressing need for rules that would help to constrain the choice of variables, and the rule we came up with—choose distinct variables to represent the states of distinct subsystems-at-times—seemed a perfectly sensible one, especially when, as in figures 7 and 8, it is perfectly obvious how to apply it unambiguously. We should expect *some* reason to think that this rule will lead us astray, and Halpern & Pearl offer none.

Consider, finally, the remarkable claim that the quoted passage ends with: "keeping this contingency in mind, it should not be too odd to name the switch position a cause of the train arrival". Well, let's test this claim. Re-read The Engineer. Focus on the following possibility: there *could* have been a flaw in the right-hand track, one that would have prevented the train from reaching its destination, had the engineer failed to flip the switch. Could have been—but *isn't*. Has it now become, in your mind, "not too odd to name the switch position a cause of the train arrival"? A possible response: "The possibility of such a flaw is not merely unrealized, but *remote*. So remote, that it is inappropriate to model the situation in such a way that recognizes this possibility. But the two-variable model *does* recognize this possibility. It is, in effect, the inappropriateness of its doing so that our causal judgments are tracking, when they stubbornly refuse to countenance the switch position as a cause of the arrival." (Hitchcock 2001 develops a similar line of argument.)

This response improves upon what Halpern and Pearl say: instead of giving a bad explanation of why we might introduce distinct variables for the distinct tracks, it gives a somewhat better explanation of why we ought *not* to: doing so allows the model to represent possibilities that are too outlandish. (Never mind that *other* choices—e.g., the choice to give **G** many more possible values—would also do so.) Unfortunately, it throws out the baby with the bathwater. Recall figure 2, our simple case of early preemption. There, what allowed the H- and HP-accounts to get the right result was *precisely* that the natural causal model can represent a certain outlandish possibility: namely, a possibility in which, even though A fires and C does not, B fails to fire. So we haven't really made any progress.

What's more, this last observation about figure 2 points to a much deeper problem. The next two examples will put it vividly on display.

§5.2 Non-existent threats

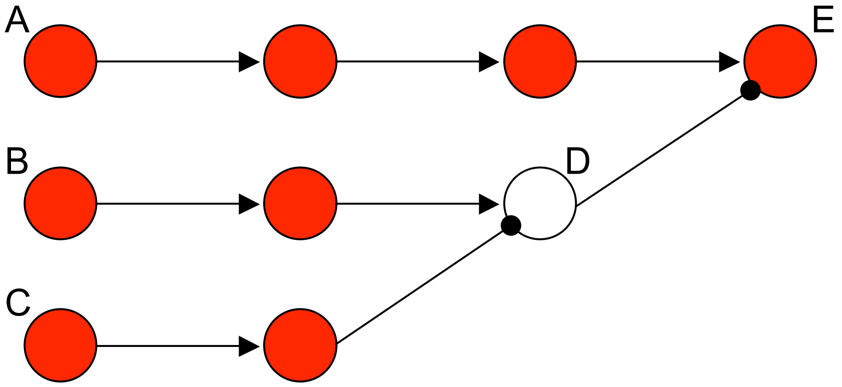


Figure 9

Figure 9 depicts a process—the one running from A to E—that is under a *threat*: for if the process initiated by B is not somehow blocked, it will end up *preventing* E. Fortunately, C fires, thus preventing the crucial intermediate neuron D from firing. E thus counterfactually depends on C, not because C is causally connected to it in a 'normal' way, but rather because C is linked to it via a two-step 'double-prevention' chain.

Let's agree, for the sake of simplifying the rest of the discussion, that C is a cause of E.[27] Certainly, the H-account, HP-account, and indeed every other structural equations account with which I am familiar will say so, since all of them take it that counterfactual dependence suffices for causation. The trouble lies elsewhere, with figure 10:

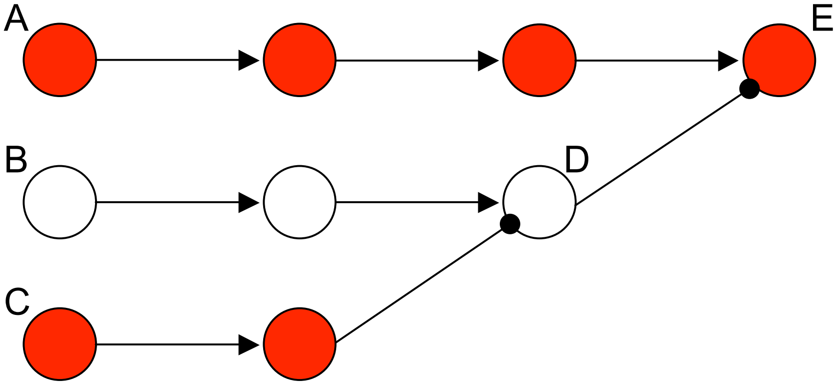


Figure 10

In figure 9, C earned the right to be counted a cause of E *solely* because it cancelled a threat to E, a threat initiated by B. In figure 10, there is no such threat. It would therefore be absurd to count C a cause of E. If you're unsure—perhaps because an excess of neuron diagrams has triggered intuition fatigue—then consider simple, real-world analogs. The family sleeps peacefully through the night, in part because the watchful police have nabbed the thief before he can enter the house. Causation, clearly. But: the family sleeps through the night, in part because the watchful police have done *nothing*, there being no thieves anywhere in the vicinity? That is plainly silly, and constitutes nothing more than a conflation of *causing* with *safeguarding*.

[27] Thus I am distancing myself somewhat from the view expressed in Hall (2004c), though largely to avoid needless complication. There, I distinguished "production" from "dependence"; what I will sketch in §7 is a kind of generalization of dependence, attractive in part because it provides a natural way to categorize certain cases of 'double prevention with backup' about which the earlier, two-fold distinction fell silent.

It is a signal failure of the HP-account (although not, interestingly, of the H-account) that it makes just this conflation. Construct the obvious causal model of figure 10, the one with these equations:

$$\mathbf{E} \Leftarrow \mathbf{A(1 - D)}$$

$$\mathbf{D} \Leftarrow \mathbf{B(1 - C)}$$

We have the actual values

$$\mathbf{A = C = E = 1}$$

$$\mathbf{B = D = 0}$$

C-D-E is a path from **C** to **E**, **B** an off-path variable. Focusing on the non-actual value **B = 1**, we have the true conditional

if (C = 0 & B = 1), then E = 0

What's more, the additional restrictive condition in the HP-account is met, as witness the true conditional

if (C = 1 & B = 1), then (C = 1 & D = 0 & E = 1)

So the model counts C as a cause of E—even when the threat C guards against is non-existent!

Notice that the arguments, such as they were, that could be called on to try to block the unwelcome result about *switches* are not available here. It's not that we can't find *some* correct causal model for the situation, according to which C does not cause E. We can *always* do that, provided that (as here) E does not counterfactually depend on C.[28] Here, for example, we could simply *omit* the variable **B** from our model. But Halpern and Pearl's insistence that "causal models earn their value in abnormal circumstances, created by structural contingencies" should, if taken seriously, force us to *include* it. What's more, what principled basis could there be for *omitting* B in this case, but *including* A and B in the model for figure 2 (essential, if the model is to count C a cause of E)? That this is what we must do, in order to get the intuitively correct result? Reading Halpern and Pearl, one often gets the queasy feeling that that is the answer working behind the scenes.

There is one more point worth emphasizing, which should quell any temptation to view the unwelcome result about figure 10 as simply displaying the "model-relativity" of causation. Such model-relativity *might* make sense if what we are doing is choosing between different *ways* of representing a given aspect of some situation. (Cases of switching strike some as a good example; see §7 for a better one.) But it doesn't make sense if, instead, what we are doing in moving from one model to another is simply *increasing* the number of aspects we are choosing to represent. Halpern

[28] Simplest way: let the model contain only the variables **C** and **E**, both treated as exogenous.

and Pearl, in other words, need to claim that, in modeling figure 10 *without* the variable **B**, we are modeling the situation *appropriately* (so: good news that this model delivers the intuitively right result that C is not a cause of E). But, they must further claim, when we simply *represent additional neurons*, we are modeling the situation *inappropriately*, and so get the wrong result. Good luck to them.

There is a lesson here about methodology. In giving an account of some philosophically interesting concept, it may be valuable to introduce a parameter that can shift with context—so that proper application of the concept is revealed to depend, in illuminating ways, on more than one might initially have suspected. When does S know that P? More carefully, under what conditions is it correct to characterize S's belief that P as knowledge? Perhaps, as recent work on contextualism about knowledge suggests, the correctness of such an attribution depends not only on whether S's belief is true, her evidence for it, and how it was generated, but *also* on the epistemic standards salient in the context in which the question of S's knowledge arises. Similarly, the truth of the claim that C is a cause of E may depend, in part, on a contextually salient way of conceptualizing the details of the situation in which C and E occur. Causal models might provide one way of making this talk of "ways of conceptualizing" precise.

Yes, this possibility is worth exploring. But one's research program degenerates if one exploits it in a cavalier manner, invoking talk of "model relativity" to block any counterexample. *Some* rules must govern such talk, and structural equations approaches have so far done a poor job of providing them.

§5.3 Short-circuits

The H-account, at least, does not fall into the trap set by figure 10. But it (hence the HP-account as well) does fall into a closely related trap:

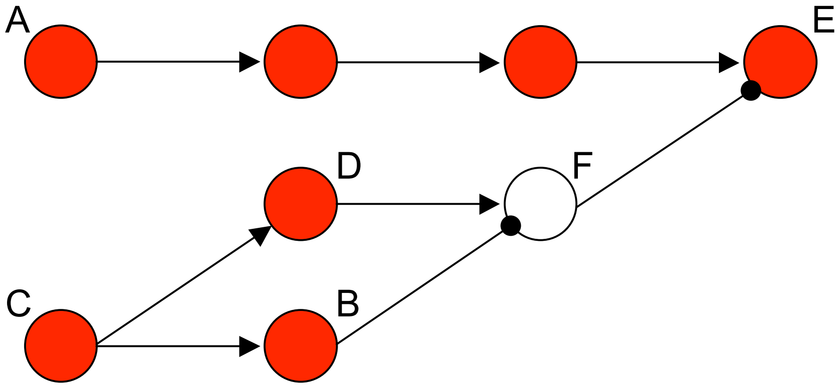


Figure 11

C initiates a threat to E: for if nothing stops D from stimulating F, then E won't fire. C also *cancels* this threat, by way of B. So the little four-neuron network C-D-B-F might aptly be called a "short-circuit", with respect to E.

I think, and most agree, that C is not a cause of E.[29] As is well known, that judgment spells trouble for the combined claims that causation is transitive, and that counterfactual dependence suffices for causation: for E depends on B, which in turn depends on C. Now, both of our structural equations accounts eschew transitivity. But that probably sensible move is of no help here, as each unavoidably counts C in figure 11 a cause of E for a *different* reason. For, *holding fixed the fact that D fires*, if C hadn't fired, then E wouldn't have.

Let's double-check this result, by constructing the obvious causal model. We have the usual binary variables $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$, $\mathbf{E}$, and $\mathbf{F}$, together with these structural equations:

$$\mathbf{E} \Leftarrow \mathbf{A(1 - F)}$$
$$\mathbf{F} \Leftarrow \mathbf{D(1 - B)}$$
$$\mathbf{D} \Leftarrow \mathbf{C}$$
$$\mathbf{B} \Leftarrow \mathbf{C}$$

The actual values are these:

$$\mathbf{A = B = C = D = E = 1}$$
$$\mathbf{F = 0}$$

Consider the conditional

$$\textbf{if (C = 0 \& D = 1), then E = 0}$$

We ignore the third equation. Having set $\mathbf{C}$ to 0 and $\mathbf{D}$ to 1, we update $\mathbf{B}$ to 0 by the fourth equation. We then update $\mathbf{F}$ to 1 by the second equation, and thus $\mathbf{E}$ to 0 by the first. So the conditional is true. So both the H-account and HP-account classify C as a cause of E. What's more, they do so for *exactly the same reason* that they count C, in figure 2, a cause of E—whence we now have good reason to doubt that they got that case right *for the right reasons*.

That's a *bad* outcome. I trust that it is, by now, sufficiently obvious that no amount of whining about 'model-relativity' will make it look any better.

§5.4 Late preemption

It would seem, judging from their treatments of switching, non-existent threats, and short-circuits, that structural equations account are not to be trusted as provid-

[29] For a rare—and strained—disagreement, see Lewis (2004).

ing a sufficient condition for causation. Time now to look at the most famous challenge to the necessity of counterfactual analyses, which is late preemption. I will rely on the classic example:

Suzy First: Suzy, an expert rock-thrower with a taste for minor acts of destruction, throws a rock at a bottle. The rock hits the bottle, shattering it. Suzy's friend Billy throws a rock at the bottle, too. He's just as expert as she is, but a bit slower. Consequently, her rock gets there first; but if she hadn't thrown it, the bottle would have shattered all the same, thanks to his throw.

It is supposed to be a major achievement of structural equations approaches that they offer a powerful new technique for dealing with such a stubborn counterexample to so many counterfactual analyses. In fact, they fail rather miserably, and at a surprising stage: the causal models standardly offered as representations of cases like Suzy First are simply *incorrect*. I'll consider one example, taken from Halpern & Pearl (2005).

Superficially, the model is quite elegant and simple; judging from various conversations I've had, it is currently thought of as providing the canonical structural equations treatment of Suzy First. It makes use of just five variables:

ST: has value 0 if Suzy does not throw; 1 if she does.

BT: has value 0 if Billy does not throw; 1 if she does.

SH: has value 0 if Suzy's rock does not hit the bottle; 1 if it does.

BH: has value 0 if Billy's rock does not hit the bottle; 1 if it does.

BS: has value 0 if the bottle does not shatter; 1 if it does.

We should understand each of these variables as making implicit reference to a particular time. More specifically, let's stipulate that Suzy throws at time 0, Billy throws at (slightly later) time 1, Suzy's rock strikes the bottle at time 2, Billy's rock *would* have struck the bottle at time 3 (i.e., if Suzy's had not already done so), and the bottle is in a shattered state at time 4. So **ST** characterizes what Suzy is doing at time 0; **BT** what Billy is doing at time 1; **SH** what Suzy's rock is doing at time 2; **BH** what Billy's rock is doing at time 3; and **BS** the state of the bottle at time 4. What we wish to write down are structural equations that show that it is Suzy's throw, and not Billy's, that causes the bottle to be in a shattered state at time 4.

This seems to be easy to do. Halpern and Pearl—and just about everyone else, as far as I can tell—find the following equations satisfactory:

$$\mathbf{BS} \Leftarrow \mathbf{BH} + \mathbf{SH} - \mathbf{BH}\bullet\mathbf{SH}$$

$$\mathbf{BH} \Leftarrow \mathbf{BT}(1 - \mathbf{SH})$$

$$\mathbf{SH} \Leftarrow \mathbf{ST}$$

Assuming these equations are correct, it's not too hard to confirm that both the H-account and the HP-account judge Suzy's throw to be a cause of the shattered state, by virtue of the conditional

if (ST = 0 & BH = 0), then BS = 0

It's also not too hard to confirm that neither account will judge Billy's throw to be a cause of the shattered state (I'll leave it as an exercise).

Success? Not so fast. Let's take a close look at what these equations mean. The third is unobjectionable: it says that Suzy's rock will hit the bottle iff she throws it. Given that we only mean to be considering four options for the temporally prior variables **BT** and **ST**—she throws/doesn't throw at just the time and with just the speed she does; he throws/doesn't throw at just the time and with just the speed he does—this equation is perfectly correct. The first equation might also seem correct: for it says merely that the bottle will be in a shattered state iff at least one of the rocks hits it. Likewise the second, which says that Billy's rock will hit the bottle iff he throws it, *and* Suzy's rock hasn't already hit it (for in that case, it won't be there for his rock to hit).

But now we should smell a rat. Look again, and closely, at the first two equations. The first strikes us as true in part because, when we envision a situation in which **BH = 0** and **SH = 0**, we understand that **BH = 0** *because Billy's rock isn't thrown*, and instead lies idle (we may suppose) in his hand. But the second strikes us as true in part because, when we envision a situation in which **BT = 1** and **SH = 1**, we understand that **BH = 0** *because the bottle isn't there to be hit.* The model is simply trading on this ambiguity in the content of the claim "**BH = 0**". Remove the ambiguity, and one or the other of the first two equations must be revised.

Let's work through this again, slowly and systematically, making explicit use of the truth-conditions for structural equations spelled out in §4. Those truth-conditions will straightforwardly vindicate the third equation. As for the second, we begin by observing that the variables that are candidates for figuring in the equation for **BH** are **ST**, **BT**, and **SH**, since these are the only variables temporally prior to **BH**. There are eight settings for these three variables to consider. Following the sequential updating version of the Maudlin-recipe, we can immediately see that in any counterfactual situation in which **ST = 1** and **SH = 1**, the bottle must be in a shattered state immediately after time 2 (so: before time 3). Forward evolution in accordance with the laws will give us a time-3 state of the world in which the bottle is *still* shattered; hence **BH = 0**, regardless of the value of **BT**—*provided* we understand "**BH = 0**" as meaning simply that Billy's rock fails to strike the bottle.

(Shortly, we'll see reasons to understand it differently.) That takes care of two of the eight cases. Suppose next that **ST = 0** and **SH = 0**. Then, clearly, **BH = 1** iff **BT = 1**. That takes care of two more cases. There are two more cases in which **BT = 0**, in both of which **ST ≠ SH**; it's not clear what goes on with Suzy's rock in those cases, but at any rate we can be sure that **BH = 0**, since Billy's rock isn't even thrown. Now to the two remaining cases:

BT = 1, **ST = 1**, **SH = 0**. Here, the time-0 state of the world is just the actual state, and no local modifications are necessary until we reach time 2, at which point we need to adjust the state so that **SH = 0**. Now, just how exactly do we do this? If the claim "**SH = 1**" is supposed to mean that Suzy's rock strikes the bottle in a certain way—namely, the way in which it *actually* strikes the bottle—and if "**SH = 0**" is supposed to be true iff "**SH = 1**" is false, then the problem is that there are far too many ways to locally modify the state so that **SH = 0**, and no principled way to choose among them. Worse: *some* of these modifications will yield forward evolutions in which the bottle is shattered, in which case we won't get the result that Halpern and Pearl apparently think is so obvious: namely, that for this choice of values, the correct equation must yield **BH = 1**. For example, Suzy's rock might strike the bottle in a rather different way from how it actually does, but still hard enough to break it.

Well, can't we just read "**SH = 0**" as saying that *Suzy's rock doesn't strike the bottle*? If so, it obviously doesn't strike it in a different way! But that isn't really any help, for we are still left with the mystery as to how to locally modify the state of the world, so as to secure the truth of this claim. More to the point, *one* way to effect this modification is to have the bottle *be in a shattered state* before Suzy's rock can strike it, whence forward evolution will give us the unwanted **BH = 0**, as before. How, exactly, are we supposed to rule out such a modification as illegitimate?

As far as I can tell, the only clean, non-ad-hoc way to secure the desired result is to read "**SH = 0**" as meaning that Suzy's rock is simply *absent* (absent, that is, from the neighborhood of the bottle—perhaps we should return it to Suzy's hand…), at the relevant time. Let us so read it. Then granted: **BH = 1**.

BT = 1, **ST = 0**, **SH = 1**. Here, the time-0 state of the world is one in which Suzy is not throwing, but, as in the actual world, Billy is preparing to throw. Forward evolve this state until time 1. Billy throws (so no local modifications to the state are necessary). Forward evolve the resulting state until time 2. Make local modifications, so that Suzy's rock—which, remember, has been sitting in her hand—hits the bottle. Once again it's not so clear how to proceed, since we haven't said with enough specificity what the content of the claim "**SH = 1**" is. So let's correct that oversight, by stipulating that this claim says that Suzy's rock hits the bottle in just the way it does in the actual situation.[30] Now we can proceed: the bottle breaks. Forward evolving, we see that Billy's rock doesn't strike the bottle. But it doesn't follow that **BH = 0**—*that* depends, unsurprisingly, on what the precise content of this claim is. If we take our cue from the foregoing discussion of "**SH = 0**", we will say that "**BH = 0**" is *false* in this situation, since Billy's rock is *not* absent from the given region: it's there all right, it's just flying over scattered shards of bottle-glass. But presumably this is not what Halpern and Pearl have in mind. So let's take it that "**BH = 0**" means simply that Billy's rock doesn't strike the bottle. Then granted: **BH = 0**.

We've now secured the second of the three structural equations, albeit at some cost: we were forced, somewhat surprisingly, to treat "**BH = 0**" as not at all analogous to "**SH = 0**". As you may have guessed, there is worse trouble ahead. We run into it as soon as we try to write down an equation for **BS**.

Consider the values **BT = 1**, **ST = 0**, **SH = 0**, **BH = 0**. What should be the corresponding value for **BS**? It should be **BS = 0**, we are told—after all, we are describing a situation in which neither rock strikes the bottle. But having been alerted by the foregoing discussion, we can easily see how this response involves some sleight-of-hand. Let's work through the case systematically, to pin down where the fallacy is lurking.

For the given values of the 'input' variables, we start with a time-0 state locally modified from the actual state, so that Suzy does not throw. Evolve forward to time 1. Billy throws (no modification necessary). Evolve forward to time 2. Suzy's rock is absent from the region of the bottle (no modification necessary). Evolve forward to time 3. Billy's rock is about to strike the bottle—and now we need to make local adjustments, so that it doesn't. Not *not* NOT so that his rock is *absent*: we know how to do that (just change the world-state to that no rock is there, replacing it by air), and we know that forward evolution of such a modified state will yield a time-4 state in which the bottle is not shattered. But all this is entirely irrelevant, since we already know, from having thought through the equation for **BH**, that "**BH = 0**" had better *not* mean that Billy's rock is absent from the region of the bottle (else that equation is simply *incorrect*); rather, it had better mean that Billy's rock—one way or another—does not strike the bottle. And with that meaning, the instructions to locally

[30] To clarify: There is a certain way that Suzy's rock hits the bottle in the actual situation, characterized by a certain velocity, approach vector, etc. "**SH = 1**" says that Suzy's rock hits the bottle in *that* way.

modify the state of the world so that $\mathbf{BH} = 0$ are simply too ambiguous: *one* way is to remove Billy's rock, but *another* way is to change the state of the bottle from whole to shattered. These different modifications yield different forward evolutions—and different values for $\mathbf{BS}$. So we cannot, after all, even *write down* a correct equation for $\mathbf{BS}$.

What is the root of the problem? Hint: you've seen such a problem before, in the faulty models of figures 5 and 6 discussed toward the end of §4. Let's take our cue from that discussion, and fix things up.

The key is to recognize that at time 3, there are *three* distinct states of affairs that we will want the variable $\mathbf{BH}$ to represent. $\mathbf{BH}$ should have one value if Billy's rock is simply absent from the region of the bottle (which is what will happen, if Billy doesn't throw). It should have another value if Billy's rock is striking the bottle (which is what will happen, if Billy throws but Suzy doesn't). And it should have a third value if Billy's rock is flying over scattered shards of bottle-glass (which is what will happen, if Billy and Suzy both throw). Let these values be 0, 1, and 2, respectively. As for $\mathbf{SH}$, we will keep it two-valued: $\mathbf{SH} = 0$ if Suzy's rock is absent from the region of the bottle; $\mathbf{SH} = 1$ if it strikes the bottle. *Now* we can apply the Maudlin-recipe cleanly; for we no longer are trying to make the single value $\mathbf{BH} = 0$ do double duty, signifying that Billy's rock is absent, when used in the equation for $\mathbf{BS}$, but that Billy's rock doesn't strike the bottle, when derived using the equation for $\mathbf{BH}$. The following equations result:

$$\mathbf{BS} \Leftarrow \mathbf{sign(SH + BH)}$$
$$\mathbf{BH} \Leftarrow \mathbf{(SH + BT)BT}$$
$$\mathbf{SH} \Leftarrow \mathbf{ST}$$

Now that we have a causal model that we need not feel ashamed of, do we at last get the right result—that Suzy's throw, and not Billy's, is a cause of the bottle's shattered state? That depends. Deploy the H-account, and you don't. Deploy the more permissive HP-account, and you do. Let's look at the details.

The only relevant path from $\mathbf{ST}$ to $\mathbf{BS}$ is $\mathbf{ST\text{-}SH\text{-}BS}$. The off-path variables have values $\mathbf{BT} = 1$ and $\mathbf{BH} = 2$. Either of two conditionals could thus testify to $\mathbf{ST}$'s causal status with respect to $\mathbf{BS}$. But both of them are false:

if (ST = 0 & BT = 1), then BS = 0

if (ST = 0 & BH = 2), then BS = 0

Suppose we had reported the actual value of $\mathbf{BH}$ this way: $\mathbf{BH} \neq 1$. Then we might think to confirm the causal standing of $\mathbf{ST}$ by means of the following conditional:

if (ST = 0 & BH ≠ 1), then BS = 0

This conditional can seem to be *true*, if what we have in mind as a situation in which **BH ≠ 1** is a situation in which the bottle remains intact, but Billy's rock somehow never reaches it. But, as soon as we take care to distinguish this situation from the situation in which Billy's rock does not strike the bottle *because the bottle is already shattered*, we expose this reasoning as fallacious. In fact, the correct causal model simply fails to assign a truth-value to this conditional. And that is because the antecedent is ambiguous. Disambiguated one way, we get the relevant but *false* conditional

if (ST = 0 & BH = 2), then BS = 0

Disambiguated the other way, we get the true but *irrelevant* conditional

if (ST = 0 & BH = 0), then BS = 0

Thus, we have made exactly zero progress on the problem of late preemption, at least within the confines of the H-account.

Intriguingly, the HP-account fares better, precisely because it allows us to find "witnessing" conditionals in which the off-path variables have non-actual values. This one will do:

if (ST = 0 & BT = 0), then BS = 0

The corrected equations count this conditional as true. What's more, this choice of value for BT meets the additional restrictive condition:

if (ST = 1 & BT = 0), then (ST = 1 & SH = 1 & BS = 1)

So far, so good: Suzy's throw gets to cause the bottle to be shattered. How about Billy's throw? The only path is **BT-BH-BS**, with values **1-2-1**. The only relevant way to set the values of off-path variables is this:

if (BT = 0 & SH = 0), then BS = 0.

But the second condition fails, since

if (BT = 1 & SH = 0), then (BT = 1 & BH = 2 & BS = 1)

is false. Success: the HP-account—when it uses the right causal model!—gets Suzy First right. But given the extreme permissiveness of the account, this success is no real cause for celebration: remember that it is exactly the flexibility to allow off-path variables to take on non-actual values that forces this account (unlike the H-account) to count as causes things that guard against non-existent threats. We'll see in §7 how to fix this problem, once we expose what is, I think, the deepest mistake committed by these structural equations accounts: their failure to incorporate any distinction between the default behavior for some bit of the world, and deviations from that behavior. That exposé comes next.

§6 The default/deviant distinction

Let us examine two different cases. They should strike you as having markedly different causal structures. The first is a minor variant on the 'short-circuit' of §5.3:

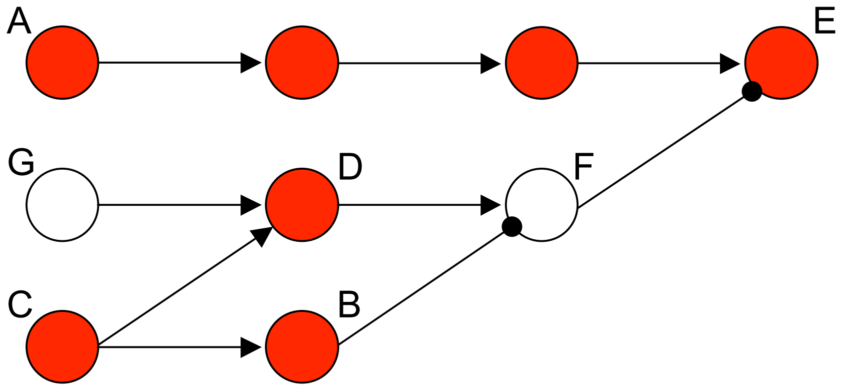


Figure 12

G does not fire. If it *had*, then E would have depended on C—for C would have cancelled not only the threat that it itself initiates, but an *independent* threat as well. In such a case, we might well count C a cause of E. But in the present case, that is a mistake. G's actual behavior poses no threat to E, so while C certainly *safeguards* E against the possible threat of G's firing, we should not conclude that C is among E's *causes*.

Maybe you don't agree. Never mind. All that really matters, for present purposes, is that we see clearly that, whatever causal structure we might wish to impute to the events in figure 12, it should be a *different* causal structure from that exhibited by the next case. To help your intuitions along, we will build up to that case in stages:

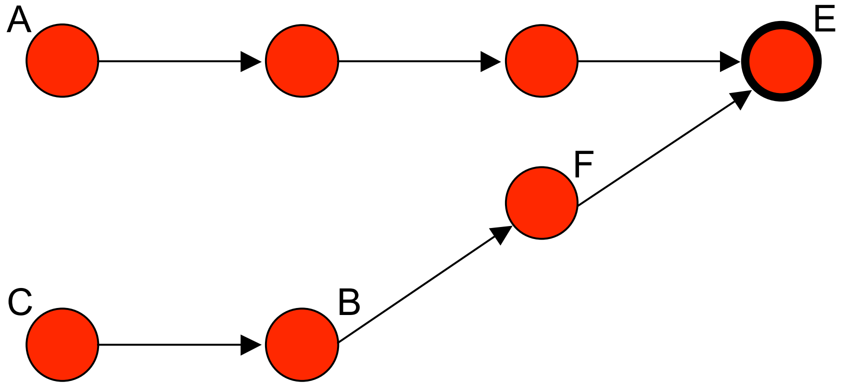


Figure 13

Neuron E in figure 13 is stubborn, needing two stimulatory signals in order to fire. It gets them: one from A, one from C. So far, the causal structure is quite clear: A and C are both causes—joint causes—of E. Now we will add a slight wrinkle:

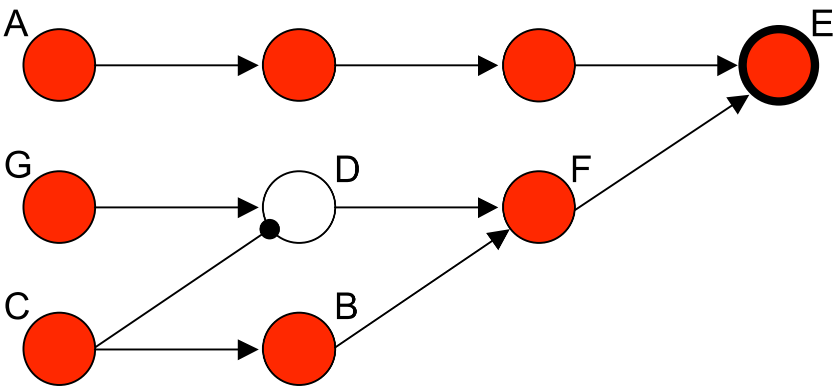


Figure 14

Look at the little network G-D-C-B-F. You've seen it before, in figure 2: it's a simple example of early preemption. We know how to think about those cases: C is a cause of F, whereas G is a preempted backup. So figure 14, although more complicated than figure 13, isn't at all hard to understand: C is a cause of F, and therefore, with A, a joint cause of E; G is not a cause of F, although it would have been, had C not fired. There is absolutely no mystery here.

Now I want you to compare figures 12 and 14. I do not ask that you agree with me about their causal characteristics; in particular, you might find figure 12 too confusing. (I rather doubt it. But in my experience, some philosophers who really ought to know better claim to be unclear about the causal structure of figure 12.) I *do* ask that you agree that their causal structures are *different*. For myself, one difference could not be more obvious: C in figure 12 is not a cause of E; C in figure 14 *is* a cause of E. But again, it's enough that you recognize that *some* difference exists. Or, even more cautiously: that a good account of causation ought to treat these two cases differently.

Why am I harping on what, I hope, strikes you as so obvious a point? Perhaps you've already spotted the reason, but anyway here it is: any structural equations approach—any *whatsoever*—that holds that the causal structure of a situation is fixed by the correct causal model or models for it, together with the actual values instantiated in it, *must treat these cases as exactly alike*. And that is because their causal models are perfectly isomorphic. Maybe that's obvious; at any rate, let's confirm it.

For figure 12, our model will include the obvious seven binary variables. Here are their equations:

$$\mathbf{E} \Leftarrow \mathbf{A(1 - F)}$$

$$\mathbf{F} \Leftarrow \mathbf{D(1 - B)}$$

$$\mathbf{D} \Leftarrow \mathbf{G + C - GC}$$

$$\mathbf{B} \Leftarrow \mathbf{C}$$

The model for figure 14 will likewise include seven binary variables, this time with these equations:

$$\mathbf{E} \Leftarrow \mathbf{AF}$$

$$\mathbf{F} \Leftarrow \mathbf{D + B - DB}$$

$$\mathbf{D} \Leftarrow \mathbf{G(1 - C)}$$

$$\mathbf{B} \Leftarrow \mathbf{C}$$

These models look different, of course. But the differences are superficial; the models are in fact the same. Remember that the numbers we use as values for our variables are *completely arbitrary*. For example, in modeling figure 12 we could decide that each binary variable has value 5 if the corresponding neuron fires (at the relevant time), and value 18 if it doesn't. The exact form of our equations will reflect these choices; for example, with the values 5 for firing and 18 for not, the first equation would need to be rewritten:

$$\mathbf{E} \Leftarrow \mathbf{18 - (18 - A)(F - 5)/13}$$

We could achieve exactly the same effect by introducing different variables, defined in terms of the original ones. (E.g., let $\mathbf{E^* =_{df} 18 - 13E}$.)

Accordingly, let's rewrite the equations in the model for figure 14, using new variables in place of $\mathbf{D}$, $\mathbf{F}$, and $\mathbf{G}$:

$$\mathbf{D^* =_{df} 1 - D}$$

$$\mathbf{F^* =_{df} 1 - F}$$

$$\mathbf{G^* =_{df} 1 - G}$$

Then—making the substitutions just on the left-hand sides—the four equations become

$$\mathbf{E} \Leftarrow \mathbf{AF}$$

$$\mathbf{F^*} \Leftarrow \mathbf{1 - D - B + DB}$$

$$\mathbf{D^*} \Leftarrow \mathbf{1 - G(1 - C)}$$

$$\mathbf{B} \Leftarrow \mathbf{C}$$

Substituting the new variables in on the right-hand sides, these become

$$\mathbf{E} \Leftarrow \mathbf{A(1 - F^*)}$$

$$\mathbf{F^* \Leftarrow D^*(1 - B)}$$
$$\mathbf{D^* \Leftarrow G^* + C - G^*C}$$
$$\mathbf{B \Leftarrow C}$$

That the two models are in fact the same is now obvious. Finally, observe that in figure 12, the actual values are these:

$$\mathbf{A = B = C = D = E = 1}$$
$$\mathbf{F = G = 0}$$

In figure 14, the actual values are these:

$$\mathbf{A = B = C = D^* = E = 1}$$
$$\mathbf{F^* = G^* = 0}$$

Suppose an account of causation tries to render a verdict about what causes what, in figure 12, making use *just* of the structural equations for figure 12, plus the actual values of the variables. Suppose that account tries to do the same, for figure 14. Then the isomorphism between the models establishes—*conclusively*—that the account will call C in figure 12 a cause of E iff it likewise calls C in figure 14 a cause of E. More generally, it will inevitably be forced to say that the two causal structures are the same. But they aren't. So something has gone badly wrong.

It should be clear what it is. The broad class of accounts we are considering (of which the H- and HP-accounts are both instances) make no provision for the possibility that what causes what might be a function, not merely of the abstract patterns of counterfactual dependence that the various states of bits of the world enter into, but also of *the intrinsic nature of those states themselves*. In figure 12, the state neuron F is in, at the relevant time, is a *non-firing* state; in figure 14, the corresponding state of F—the state that occupies the same location within the abstract structure of counterfactual dependencies—is a *firing* state. It must be this difference (and the corresponding difference in the states of D and G) that matters.

Well, what *is* this difference? That is, what sort of general characterization ought we to give of it? This one, I suggest: it is the difference between a *default* state of a system and a *deviation* therefrom. Neurons can be in various different states: they can be dormant; they can fire this way; they can fire that way; and so on. There is a natural distinction to draw between these states: dormant on one side, all the rest on the other. Likewise, bottles can be in various states: they can lie there placidly, or shatter, or melt, and so on. Again, there is a natural distinction to draw: lying there placidly on one side, shattering et. al. on the other. And so it goes: we very often find, in contemplating various parts of the world, that we have a reasonably clear and firm conception of what that part would be doing if nothing was happening to it. That is its default state; anything else counts as a deviation.

That test—a system's default behavior at a time is the behavior it would exhibit, were nothing acting on it—is explicitly causal, thanks to the word "acting". At this point, I do not know whether we can provide a fully general test that *isn't* causal, tacitly or explicitly. In *certain* kinds of cases we can provide a test: for sometimes we can pick out, in a sufficiently precise and non-arbitrary manner, a state of the system's *environment* that qualifies as a state in which *nothing is happening*—a fortiori, a state in which nothing is *acting on* the system. Example (borrowed from Maudlin's discussion in his 2004): A Newtonian particle will exhibit a certain distinguished behavior—constant, linear motion—in an environment in which nothing else exists. (That's a fancy way of saying: in a world with just that particle in it, it will move with constant velocity.) Obviously, if nothing else exists, then there is nothing else that can be the subject of a "happening". So we have our test environment, non-causally characterized, and can use it to define default behavior for a Newtonian particle: it is just constant linear motion. Maudlin makes a persuasive case that Newton's First Law—which, from a mathematical standpoint, is perfectly redundant, being a trivial consequence of the Second Law—in fact plays an important expository role, precisely because it explicitly articulates the default behavior for a Newtonian object.

Alas, I think it is simply not to be hoped that, for every case in which there is clear agreement about the default/deviant distinction, the default behavior can be analyzed as the behavior the system in question would exhibit, if it were in an environment in which nothing else was happening. Consider people, whose default physiological behavior is to go on *living*. That seems right, at least as *one* legitimate way to draw the default/deviant distinction in this case. (Complications will appear shortly, but don't matter for the present point.) But living is precisely *not* what they would continue to do, if they were in an environment devoid of happenings (let alone in an environment in which nothing else existed!).

Perhaps another test will work: the default behavior for a system, at a time, is the behavior it would exhibit if it did not undergo any *change*, at that time. *Any* change? Then the living person who goes on living is deviating from default behavior. Well, maybe no change, outside of certain difficult-to-specify parameters. But even if that test worked, it would not cover the case of the Newtonian particle—nor the case of any system whose default behavior is *periodic* (think of a pendulum, as a canonical example).

So a comprehensive, illuminating account of the default/deviant distinction is not going to be easy to find. Never mind; we can leave the search for it for another day. What I mainly wish to demonstrate, in what follows, is that the distinction provides the key to a very simple and attractive account of causation. It will be enough that we agree, in particular cases, on how to *draw* the distinction. I will try to help foster such agreement with a few observations. They fall regrettably short of anything like a proper theory!

First, we have already encountered one important role for the default/deviant distinction, in evaluating counterfactuals that concern what would have happened, had some (actual) event not occurred. A conditional of that form—"if event C had not occurred, then…"—has a highly non-specific antecedent. Even if we agree that the counterfactual situation described is one in which the *rest* of the world, apart from that bit of it that is involved in C's occurrence, is in the same state as it *actually* is at the time in question, there are indefinitely many ways to fill in the remaining details. You walk into a room, and flip a switch, turning on (or off; it doesn't matter which) the lights. What would have happened if that switch-flipping hadn't occurred? More obviously, what would have happened if you hadn't flipped the switch? The question does not direct our attention to a situation whose character, as regards the switch's behavior, is highly indeterminate; we know perfectly well that we mean to be talking about a counterfactual state of the world in which the switch's position remains unchanged. Or, as I would put it: a counterfactual situation in which the switch is in its default state.

It is worth noting the contrast between the ease with which we evaluate this counterfactual, and the difficulty we find in evaluating counterfactuals of the same form, but that concern systems for which the non-arbitrary assignment of a default state is impossible. As an artificial but particularly vivid example, consider a cellular automaton in which each cell can have, at each moment, one of four different colors: red, blue, green, and yellow. A deterministic rule fixes the state of each cell at time t+1 as a function of the state of it and its eight neighbors at time t. This rule, furthermore, fails to distinguish any of these states as in any sense a dynamically "inert", or "nothing happening" state.[31] Accordingly, there is no sense in trying to figure out what a cell's state would be, at a given time, if nothing were happening to it: for the laws of this little universe guarantee, as it were, that something is *always* happening to *every* cell.

Let event C consist in a particular cell A's being red, at a particular time t. If we ask, "What would have happened, had C not occurred?", we do not construct a *single* counterfactual t-state; rather, we construct *three*, by holding the state of every other

[31] How might the dynamics distinguish one state as a 'nothing happening' state? Perhaps this way: there might be a unique state such that, if *every* cell has that state at some time, then given the dynamics, every cell must *continue* to have that state, thereafter. The idea is that the characteristic dynamical behavior of a state of the world that qualifies as a state in which nothing is happening, anywhere, is to persist unchanged. Note that in Conway's game of "Life", the 'empty' cell state has this feature, but the 'filled' state doesn't.

cell fixed and letting cell A be green, blue, and yellow, respectively. What would have happened is exactly what the cell-dynamics entail, regardless of which of these three states we choose. Lacking a default state to 'return' cell A to, we exercise the only other option: let A counterfactually run through *every* available state that is compatible with our antecedent. If you need a reminder of how pervasive the default/deviant distinction is in our everyday counterfactual reasoning, you need only reflect how rare it is to find a real-world analog of this example.

Second, the default state of a system can change with its circumstances. If a bottle is intact, its default behavior is (among other things) to remain intact; if it is shattered, default behavior is to remain shattered. Similarly with people: if alive, dying counts as a deviation; if dead, resurrection likewise counts as a deviation. (Here I'm especially indebted to some cogent observations of Chris Hitchcock's.) Not so with our neurons: the default state for a neuron, at any time, is to be dormant. But that was a byproduct of optional stipulations. Suppose we modify those stipulations, so that neurons are like switches: then, if switched on, their default behavior is to stay on; if switched off, to stay off.

Third, what counts as a default state is not, I think, a purely objective matter. (Well, maybe it is in some cases: e.g., for Newtonian particles.) Context can, within severe constraints, affect what counts as the appropriate default state for some part of the world. Example: A large rock sits in a sealed room, at noon. Arrayed around the room are sensitive detectors, which will trigger an alarm if they register a sudden pressure change in the room. We ask: what would have happened, at noon, had the rock not been present? That is, what would have happened, had there been no rock in the region of the room where there is in fact a rock? Two contradictory answers are available—each defensible, because each makes tacit use of a different but equally legitimate choice of default state, for that region of the room. First answer: nothing would have happened; so the presence of the rock makes no difference to whether the detectors trigger the alarm. Second answer: without the rock there, a sudden drop in pressure would ensue, as air rushed to fill the empty space; so the presence of the rock is helping to *prevent* the detectors from triggering the alarm. You might find one answer more persuasive than the other. But I think, in fact, that any attempt to rank them is a mistake, which can be brought out by considering this question: What is an appropriate default state for the given region of the room? –A state in which nothing occupies it, one is tempted to answer. That invites a follow-up: Nothing *at all*, or just nothing but what would *normally* occupy it (viz., *air*)? Choose the first answer, and you will judge that without the rock, there would be a sudden drop in pressure; choose the second, and you'll deny this claim. But there is no real conflict here—just a difference between equally acceptable ways of filling in the details of the counterfactual situation that we specify indeterminately as one in which the rock is absent.

The example reveals not only a context-sensitivity in the default/deviant distinction, but a way in which that sensitivity can influence *causal* judgments: whether or not we judge the presence of the rock to be preventing the alarm from going off depends on what we take to be the given region's default state. We'll see a subtler example of the same phenomenon in the next section. To begin, though, we'll stick to easier cases, where the relevant default/deviant distinctions are clear and unambiguous. Writing these distinctions into our account of causation makes it surprisingly easy to give a uniform treatment of the sorts of cases that spelled trouble in §5.

§7 An improved account

The account I will offer makes use of the following guiding idea: What causes what is a matter of the intrinsic character and relations among the events involved. As always with guiding ideas, this one can motivate different proposals, differing in crucial details. I used to think that the right proposal would need to rest on the following thesis, which I viewed as a more precise statement of the guiding idea (Hall 2004a):

Intrinsicness: Let S be a structure of events consisting of event E, together with all of its causes back to some earlier time t. Let S' be a structure of events that intrinsically matches S in relevant respects, and that exists in a world with the same laws. Let E' be the event in S' that corresponds to E in S. Let C be some event in S distinct from E, and let C' be the event in S' that corresponds to C. Then C' is a cause of E'.

I used to think that Intrinsicness provided the key to one paradigmatic kind of causation—what I called "production"—in which the cause brings about its effect by way of a connecting process. Production, I thought, should be contrasted with *dependence*, a more minimal kind of causation in which the only connection between cause and effect is that the latter counterfactually depends on the former. I had hoped for a simple 'two concepts' story, according to which production and dependence typically go hand in hand, but can sometimes come apart: thus, Suzy First would be a paradigm example of production without dependence, cases of threat-canceling dependence without production. Production, finally, struck me as in some sense the more important notion, though not so important as to make dependence irrelevant to a proper account of causation:

> A third, more congenial objection begins by granting the distinction between production and dependence, but denies that dependence deserves to be counted a kind of causation at all. Now, I think there is *something* right about this objection, in that

> production does seem, in some sense, to be the more "central" causal notion. As evidence, consider that when presented with a paradigm case of production without dependence—as in, say, the story of Suzy, Billy, and the broken bottle—we unhesitatingly classify the producer as a cause; whereas when presented with a paradigm case of dependence without production … our intuitions (well, those of some of us, anyway) about whether a genuine causal relation is manifested are shakier. Fair enough. But I think it goes too far to deny that counterfactual dependence between wholly distinct events is not a kind of *causal* relation. Partly this is because dependence plays the appropriate sort of roles in, for example, explanation and decision. (See §8, below, for more discussion of this point.) And partly it is because I do not see how to accommodate causation of and by omissions (as we should) as a species of production; counterfactual dependence seems the only appropriate causal relation for such "negative events" to stand in. (Hall 2004c, p. REF)

That would have been a nice story, one according to which "cause" functions like other terms for which we can articulate more than one precise account of their application conditions, accounts that typically coincide but occasionally conflict (or at least *could* do so): think of "child", or "mother".[32] The analysis of *production* articulates one set of application conditions; the analysis of *dependence* another. Or, to put the point in a mode that I prefer, production and dependence are two metaphysically distinct relations that events (and in the case of dependence, facts) can bear to each other, each of which deserves to be called "causal"; the business of the metaphysician is to explain their structure, and investigate what interesting work they can do. We can leave it to the semanticist to explain how, precisely, they connect up to our messy term "cause".

Much of that picture still strikes me as correct; in particular, I think it is useful and important to distinguish production from dependence, and to give a theory of each relation. (Production is hard; dependence is comparatively easy, being just, well, counterfactual dependence.) But two problems remain. Cases of switching pose the first problem: for Intrinsicness, plus one unproblematic assumption, guarantees that switches are causes. Recall The Engineer, discussed in §5.1. Imagine a variant, in which there simply *is no right-hand track*. Then the engineer's action unquestionably helps get the train to its destination—i.e., counts as a cause of the arrival. (That is the unproblematic assumption.) But the original case contains, we may suppose, a perfect duplicate of the events that unfold in this variant. Apply Intrinsicness, and you get the unwelcome result that even in the original case, the engineer's action is a cause of the arrival. The generalized result cannot be avoided: an account of production that rests on Intrinsicness must call switches *producers* of the relevant effect, and so in one central sense *causes*.

[32] "Child": we can focus on chronological age, or on physiology. "Mother": to be a biological mother is not the same thing as to be an adoptive mother.

The second problem arises from variants on threat-cancelling, in which a *backup* threat-canceller is present, but remains idle. Figure 15 illustrates:

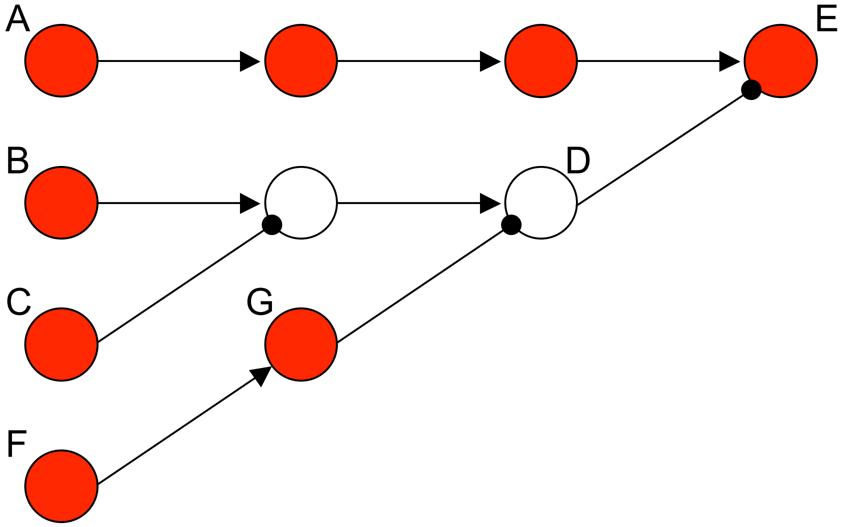


Figure 15

E faces a threat from the firing of B. C cancels this threat. But F (by way of G) would have done so, had C not occurred. E does not depend on C; nor is C connected up to E via the sort of process that would make C counts as a producer of E.[33] Given my earlier, dual-concept view, C *in no sense* counts as a cause of E. That seems wrong: F notwithstanding, it is C that *in fact* cancels the threat to E, and canceling a threat is one way to be a cause.

These are not fatal objections, particularly given the methodological stance outline in §1: I could say, I suppose, that our ordinary intuitions about cases have guided us to two potentially useful concepts, and leave it at that. But I think I can do better. There is another, subtly different way to exploit the guiding idea that what causes what is a matter of the intrinsic character and relations among the events involved. It was suggested to me by Joshua Haas[34]; I'll now try to explain it.

Imagine a situation where all sorts of things are happening. C occurs. A bit later, E occurs. Lots else occurs, besides. E does not depend on C, let's suppose. Nevertheless, it might be that the right sort of structure is in place to support such depend-

[33] That's generally true of threat-cancelers: since the presence of the threat is typically extrinsic to any reasonable candidate for a sequence of causes connecting the threat-canceler to the effect, Intrinsicness will rule that they are not causes, at least of the sort that thesis aims to characterize.

[34] In a homework exercise for my causation seminar at Harvard, spring 2006.

ence, but that events extraneous to this structure are, by their occurrence, *masking* this dependence. We can test for such masking by seeking a variant of this situation—a nomologically possible variant—in which *strictly fewer* events occur, and in which E *does* depend on C. (I.e., C and E still both occur; but if C hadn't, E would not have.) If so, then C is a cause of E: for the existence of this variant demonstrates that the underlying dependence of E on C is simply being masked.

By "strictly fewer" I mean this: that every event that occurs in the variant situation occurs in the actual situation, but not conversely. Without this rider, the test collapses, saying that C is a cause of E if there is some other situation in which E depends on C. That test is far too easy to pass. Our test isn't. The intuitive idea behind it is this: The given situation unfolds in a certain way. It could have unfolded in any of a number of different ways. In some of these ways, E depends on C. When it does so, that might be because of the presence of *novel structure*, structure constituted by events that do not occur in the actual situation. Alternatively, it might be because of the *lack* of structures that are actually present. It is this last possibility that concerns us: it shows, as it were, that C and E are "poised" to exhibit dependence, but for the masking effect of some extraneous structure of events. Such "masked dependence" suffices for causation. More generally, I propose a necessary and sufficient condition: C is a cause of E iff C and E both occur, and there is a nomologically possible situation in which (i) every event that occurs also occurs in the actual situation; (ii) E depends on C. Special case: this situation simply *is* the actual situation, whence we get the limiting result that counterfactual dependence suffices for causation.

Shortly, we'll see the need for further qualifications. But first we need to understand this talk of "situations", and of "removing" events, in a way that doesn't replace them with any new event. As for "situation", I think there will be no harm in taking a situation to consist of the entire history of the world from the time of C's occurrence to the time of E's occurrence. In practice, we'll ignore most of this history; in particular, our causal models of "situations" will be vastly more selective. That's fine, provided that the verdict about what causes what won't change, as more of the C-E history is explicitly taken into account. That condition, as we will see, sets natural limits on how selective our causal models can be.

As for "removing", what we need to appeal to is, not surprisingly, the default/deviant distinction. In one situation, lots of events occur—*that is*, various bits of the world exhibit *deviations from their default states*. In another situation, strictly fewer events occur—*that is*, some of the bits of the world that are in deviant states in the first situation are in their *default states* instead; and every other bit is in the same state as it was. (*Exactly* the same state? No. More on this qualification, later.) That is what it is for one situation to be, as I will call it, a *reduction* of another. Letting the "null" reduction of a situation just *be* that situation, we can now say the C causes E iff there is some reduction of the C-E situation in which E depends on C.

Let's consider how to implement this analysis, within the structural equations framework. We will stick with our easy neuron diagrams.[35] The key move is to require that one of the possible values for each variable be a *default value*—i.e., a value corresponding to a state of affairs in which the system characterized by that variable has its default state at the time the variable concerns. We've already met this requirement: the conventional value 0, for non-firing, will be the default value for each variable. Any other value will be a *deviant value*.

Suppose we have a causal model for some situation. The model consists of some equations, plus a specification of the actual values of the variables. Those values tell us how the situation *actually* unfolds. But the same system of equations can also represent *nomologically possible variants*: just change the values of one or more exogenous variables, and update the rest in accordance with the equations. A good model will thus be able to represent a range of variations on the actual situation. Some of these variations will be—or more accurately, will be modeled as—*reductions* of the actual situation, in that every variable will either have its actual value or its default value. Suppose the model has variables for events C and E. Consider the conditional

if C = 0, then E = 0

This conditional may be true; if so, C is a cause of E. Suppose instead that it is false. Then C is a cause of E iff there is a reduction of the actual situation according to which C and E still occur, and in which this conditional is true.

Let's put this idea into practice; along the way, we'll see why, and in what sense, an adequate causal model must be sufficiently comprehensive. Return to figure 2:

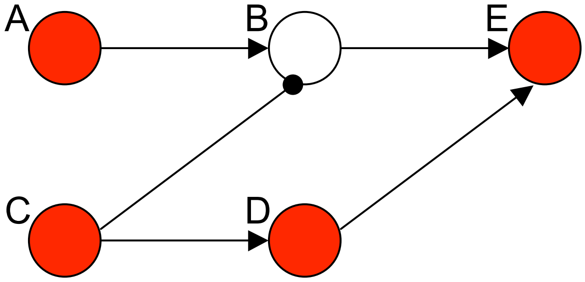


Figure 2

[35] What makes them so easy is in part that the default state—namely, non-firing—for a neuron is so clear and unambiguous, in part that this choice of default state is *fixed*, independently of its setting or history, and in part that there are so few deviations to keep track of. Remove any of these simplifying conditions, and the account inevitably becomes more complicated.

Construct the obvious causal model. According to it, the conditional

if C = 0, then E = 0

is false. But there is a reduction in which this conditional is true: namely, the variant we arrive at by setting the exogenous variables to the values **A = 0**, **C = 1**. So C is a cause of E.

Observve that A is *not* likewise a cause of E. The conditional

if A = 0, then E = 0

is false. The only variant in which it is true is the one in which **A = 1** and **C = 0**. But this is not a *reduction* of the actual situation: for **B** has the value 1, which is neither its default value nor its actual value.

Before turning to harder cases, let's stop to make an observation about good modeling practice. We could, of course, construct a three-variable causal model for figure 2, by omitting the variables **B** and **D**. Our one equation would then be

$$\mathbf{E \Leftarrow A + C - AC}$$

According to this model, both A and C are causes of E. No surprise: this model effectively (mis)treats figure 2 as a case of symmetric overdetermination. Now, we already knew that this was a bad model for figure 2. But now we can say about more about *why* it is bad. *According to the model*, the situation in which **A = 1** and **C = 0** is a *reduction* of the actual situation—since, after all, every variable *in the model* has either its actual or its default value. But this situation is, of course, *not* a reduction of the actual situation. A proper model should have recognized that fact. So a hard and fast constraint emerges on models: an adequate model must include enough variables and values that it does not represent a variation on the actual situation as being a reduction, when it is not.

Let's cruise now through the problem cases. First, switches:

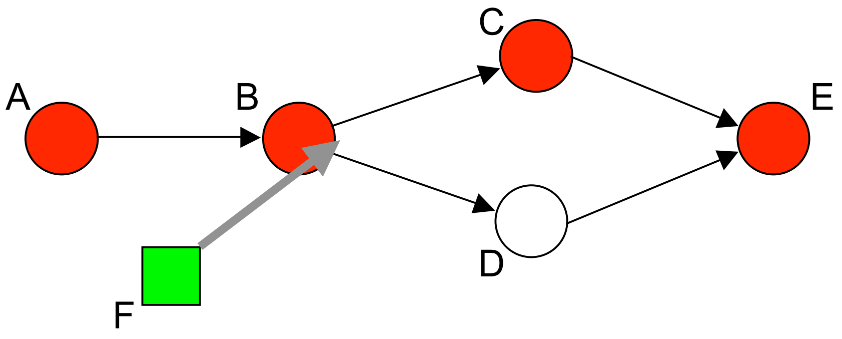


Figure 7

Here are the equations:

$$\mathbf{E} \Leftarrow \mathbf{C + D - CD}$$
$$\mathbf{D} \Leftarrow \mathbf{B(1 - F)}$$
$$\mathbf{C} \Leftarrow \mathbf{BF}$$
$$\mathbf{B} \Leftarrow \mathbf{A}$$

The variables have these actual values:

$$\mathbf{A = B = C = F = E = 1}$$
$$\mathbf{D = 0}$$

A and **F** are the sole exogenous variables. To find a reduction in which E depends on F, we must of course let $\mathbf{F = 1}$. Then the only variation we can construct is the one in which $\mathbf{A = 0}$ and $\mathbf{F = 1}$. But then $\mathbf{E = 0}$. So the model does not even yield a *variation* in which E depends on F, let alone a reduction in which it does so. So F is not a cause of E. Note how effortlessly the approach on offer secures this result: in particular, there is absolutely no need to fret about whether the behavior of C and D should be collapsed into a single variable.

Next, non-existent threats. Here we need not even bother reproducing the example; a casual glance back at figure 10 will confirm that there is no reduction in which E depends on C, and more generally that no event will count as a cause simply because it offers safeguards against a non-existent threat.

Next, short-circuits. As with switches, the verdict comes quickly: there is no variant in which E depends on C, hence no reduction in which it does so. Again, we achieve in a direct and natural way the result that C cannot be a cause of E merely by initiating a threat to E that C itself cancels.

Next, let us compare figure 12 and 14; we won't stop to reproduce the causal models. Figure 12 has one variant in which E depends on C:

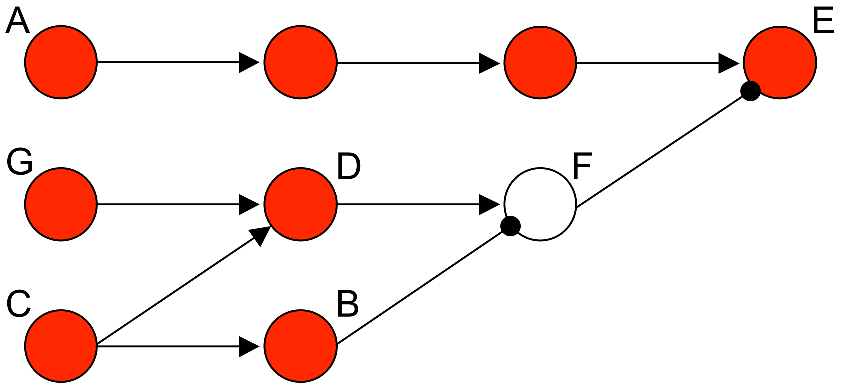


Figure 16

But this variant is not a *reduction*, since G is in neither its default state nor its actual state. Figure 14, by contrast, *does* have a reduction in which E depends on C:

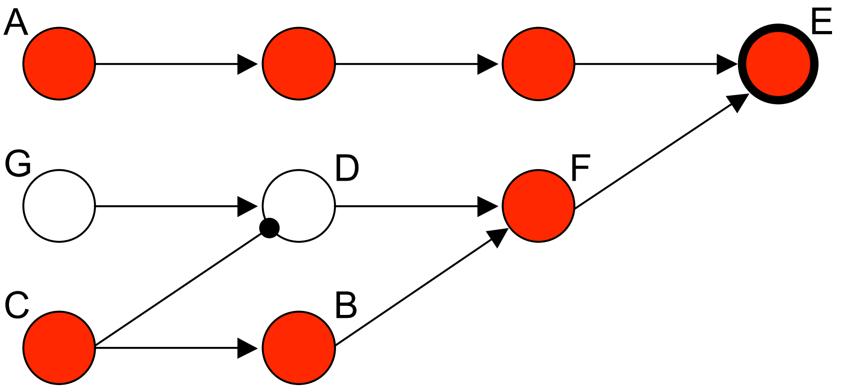


Figure 17

The account thus secures—again, rather effortlessly—the obvious contrast between the causal structures of figures 12 and 14.

Next, threat-canceling with backup. Again, the contrast is easy to see. In figure 18, we have a reduction of the situation depicted in figure 15, and E depends on C:

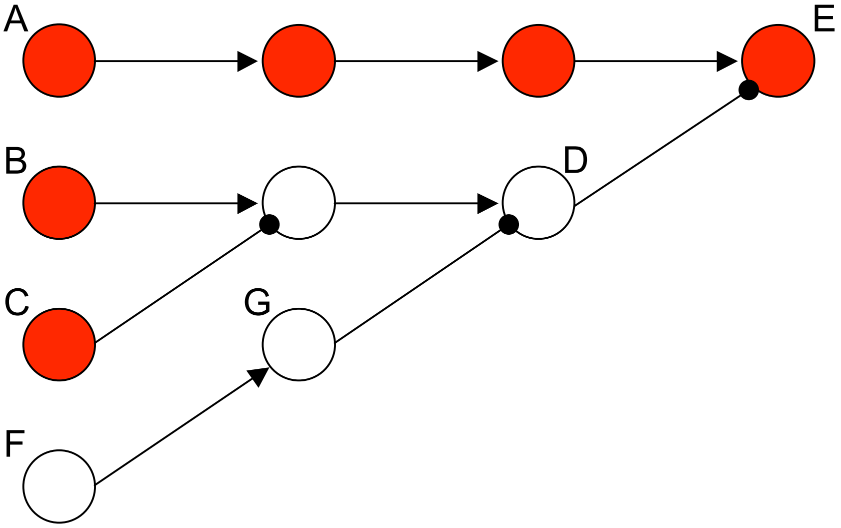


Figure 18

However, the closest we can get to a reduction in which E depends on F is this:

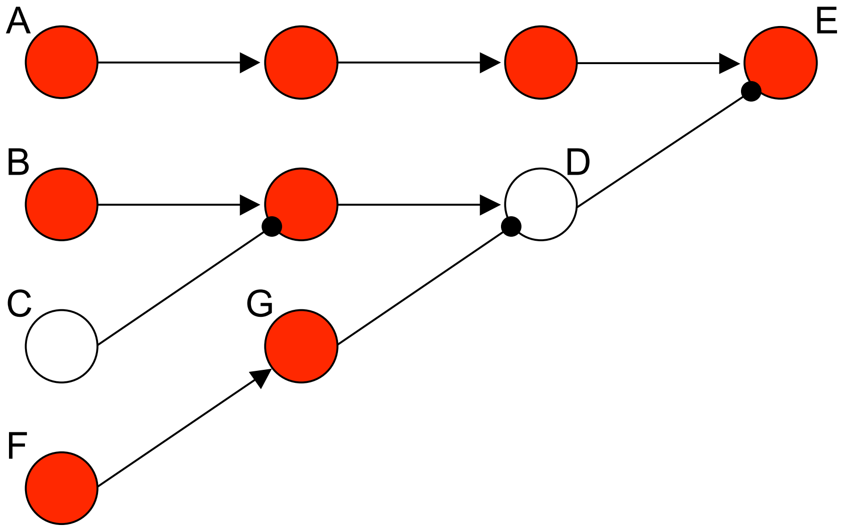


Figure 19

Not close enough, as is evident.

Finally, late preemption. Here, we can simply endorse the treatment provided by the HP-account, but with an important proviso. The HP-account, when applied to the *correct* causal model for Suzy First, said that Suzy's throw counts as a cause of the bottle's later shattered state because of the truth of these conditionals:

if (ST = 0 and BT = 0), then BS = 0

if (ST = 1 and BT = 0), then (ST = 1 & SH = 1 and BH = 1)

According to the present account, it is appropriate to focus on these conditionals—but not for the reasons the HP-account gives. We want two things: First, a reduction of the actual situation; the second conditional in effect tells us that that is what we get, when we set the values of the two exogenous variables to **BT = 0** and **ST = 1**. And second, we want the shattered state to depend on Suzy's throw, in this reduced situation; the first conditional describes just such dependence. So the HP-account arrives at the right answer, but via a mistaken view about what causation consists in.

We can go a bit further. Suppose that C is a cause of E, according to the proposed account. Then there is some causal model of the situation in which C and E occur, that has the following features: There are zero of more exogenous variables $\mathbf{X_1}$, …, $\mathbf{X_n}$, such that the situation in which they all have their default values (i) is one in which, given the equations in the model, all variables have either their default or their actual values; (ii) is one in which $\mathbf{C}$ and $\mathbf{E}$ have their actual values; and (iii) is in one that, according to the model, makes the conditional

$$\mathbf{if\ C = 0,\ then\ E = 0}$$

true. Suppose, in addition, that there is at least one path from $\mathbf{C}$ to $\mathbf{E}$ such that, in this reduced situation, every variable on the path has its actual value. (This will typically if not invariably be the case, it being quite difficult to construct a case in which E depends on C, even though every "route" from C to E goes through default states!) Letting $\mathbf{P}$ describe these values, that means that $\mathbf{P}$ is true in the given "reduced model".

Let the actual value of $\mathbf{C}$ be $\mathbf{v}$. Let each variable's default value be 0. Then, since $\mathbf{P}$ is true in the reduced model, the conditional

$$\mathbf{if\ (C = v\ \&\ X_1 = 0\ \&\ \ldots\ \&\ X_n = 0),\ then\ P}$$

is true in the model of the actual situation, since the antecedent effectively *picks out* the reduced model. Finally, the conditional

$$\mathbf{if\ (C = 0\ \&\ X_1 = 0\ \&\ \ldots\ \&\ X_n = 0),\ then\ E = 0}$$

is likewise true in the model of the actual situation. It follows that the HP-account will classify C as a cause of E. This result lets us see quite precisely a way in which the HP-account was a step in the right direction. It misfires, not because it allows off-path variables to have non-actual values, but because it badly mischaracterizes the additional constraints that need to be imposed.

The case of Suzy First brings up an additional issue, one that remains invisible when we focus on the highly sanitized neuron diagrams. The reduction of Suzy First in which the bottle's shattered state depends on Suzy's throw is a situation in which Billy doesn't throw, but rather stands there idle. It counts as a reduction in part because Suzy, her rock, and the bottle exhibit their *actual* states, at each time. Put another way, we remove some events, without changing the nature of the remaining events. Now, that characterization contains an inaccuracy: for *of course* the sequence of events leading from Suzy's throw to the bottle's shattered state will unfold in a *slightly* different manner, in the situation in which Billy doesn't throw. (The air currents are a tiny bit different, the gravitational forces on Suzy's rock are a tiny bit different, etc.) Dealing with fictional neurons, we could stipulate that the events that remain unchanged, in a reduced situation, remain *strictly* unchanged. But the real world doesn't work that way. Rather, what we need to say, somehow, is that a reduction of a situation is a situation in which some events that actually occur don't occur (being replaced by default states), and in which the remaining events do not differ in the way they occur *in any but irrelevant respects*. That's a familiar problem, faced, for example, by the attempt to treat cases of late preemption by appeal to the Intrinsicness thesis discussed above. (See Hall 2004a.) I am simply going to leave it as unfinished business to investigate whether the treatment of the problem that works for that case can be extended to the "reduction" account being sketched here.

That completes our tour of the troublesome cases. I'll consider one more case, with a rather different character: far from being a case about which intuitions are firm, it is a case about which they can flip-flop in an intriguingly systematic matter[36]:

Unprotected Window: Suzy throws a baseball at a window. Billy leaps up and intercepts it. If he hadn't, the ball would have struck the window, breaking it.

Protected Window: As before, except this time the window is protected by a high, thick wall, so that if Billy had not intercepted the ball, the ball would have bounced harmlessly off the wall.

Intuitions about Unprotected Window are clear: Billy's action prevents the window from being broken. About Protected Window, by contrast, intuition seems to equivocate. One argument, borrowed from McDermott (1995), purports to show that Billy's action *does* prevent the window from breaking: for between Billy and the wall, *something* had to stop the ball, and since the wall didn't do anything, Billy gets the credit. A second argument purports to show that what Billy did was quite inconsequential: for given the presence of the wall, the window was never under any threat. Which of these is correct?

We can make a case that *both* are correct, if we pay attention to the way the default/deviant distinction can function in this example. Here is one way: we focus on the wall as one of the components of the situation; noting that it remains in its default state at all times, we observe that there is no reduction of this situation in which the intact state of the window depends on Billy's action.[37] But here is another way: instead of focusing on the wall, we focus on the region in which it is located, and we assign to that *region* the default state *unoccupied*. Then there *is* a reduction in which the window's intact state depends on Billy's action. Now, this second choice is hardly the most natural choice—which explains, I think, why the McDermott argument has a whiff of sophistry about it. But it's not wholly illegitimate, either. (It's not as if we'd

[36] Maudlin (2004) considers very similar examples; my treatment owes much to his discussion, even though it takes a somewhat different approach.

[37] Some care is required here, in handling the notion of "reduction". Imagine a situation where the wall starts out *collapsed*. Then its default behavior is to *remain* collapsed. (Compare our earlier discussion of being alive and being dead, in §6.) Then on one reading, every component has either its actual state or its default state: Suzy, the baseball, Billy, and the window all have, at each time, their actual states; the wall has, at each time, its default state. That's not the intended reading. Rather, what we need for a reduction is a situation in which every component has, at each time, either its actual state, or *what in actuality is* its default state. Note that this problem doesn't arise if, as with neurons, a system's default state is independent of its circumstances.

said something off-the-wall like "The default behavior for the wall is to melt, and remain melted.") I suggest that what McDermott's argument does, for those inclined to treat it charitably, is to introduce a context where the contrast between a wall-present situation and a wall-absent situation becomes salient.

On the present approach, it should come as no surprise not merely that Protected Window can evoke such different reactions, but also that varying the details can make it easier to draw out one reaction or the other. As the case stands, it seems odd, even in the face of McDermott's argument, to describe it by saying that Billy's action prevents the window from breaking. But suppose we replace the wall by a backup catcher—Sally, say—poised to intercept the baseball if Billy doesn't? Let's stipulate that Sally is perfectly reliable: there is absolutely no question that she will intercept the baseball, if Billy doesn't. Then we can argue as before that the window is under no threat, hence Billy does not prevent it from breaking. But that argument does not seem quite as persuasive. The counterargument that, between Billy and Sally, someone had to stop it—and Sally just stood there idly—seems to have much more force. Why is that?

Here is one explanation. In the actual situation, Sally is alert, poised to intercept if Billy doesn't. We might focus on the contrast between that psychological state, and a state she easily could have been in, in which she not only is idle but is disposed to remain so, perceived threats to the window notwithstanding. With that as our choice of default state for Sally, a reduction is easy to find. Such a choice is reasonably natural—more natural, at any rate, than shifting attention from the wall to the region it occupies, and conceiving of the default state for that region to be unoccupied.

This completes my sketch of the "reduction" account of causation. It is just a sketch: once removed from the safe world of neuron diagrams, it faces a host of complications, best pursued on another occasion. For present purposes, I wish to emphasize two points. First, while structural equations accounts of causation are possible that improve dramatically on the poor offerings found in the literature, there is no good reason to think that causation *should* be analyzed by such means; the inflated reputation such approaches currently enjoy is due for a correction. The second point is more important: whatever the merits or defects of the "reduction" account, the ease with which it provides uniform treatments of cases as diverse as late preemption, switching, and threat-canceling with backup is too striking to be ignored. We knew that ordinary counterfactual reasoning employs the default/deviant distinction (or something like it); what the successes of the "reduction" account suggest is that this distinction operates in an even more pervasive manner in our causal reasoning. I'll close with an overview of some of the questions about this distinction that strike me as most worth investigating.

§8 Some larger questions

First, what makes the distinction tick? In §6, I offered some sketchy remark on this topic, but it's obvious that a proper theory would be welcome. I suggest that in pursuing such a theory, a good place to start is with these questions: In how many cases can the default behavior of a system be usefully defined as the behavior that system would exhibit, in a suitably canonical environment? And when it can, what is the proper characterization of this canonical environment? Here it is helpful to remember the example of Newtonian particles: the default behavior of such a particle is quite naturally picked out in this way, with the obvious choice of "canonical environment" being an environment in which that particle is the only thing that exists. One topic that bears investigation is the extent to which this example can be generalized.

Second, how does the default/deviant distinction function, in causal reasoning? The "reduction" account gives one answer, but it is important to recognize that even if that answer succeeds, it is only partial. Consider figures 5 and 6 again:

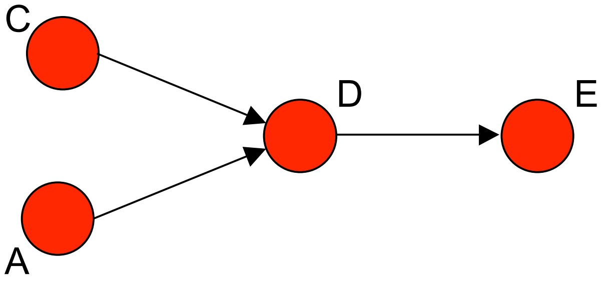


Figure 5

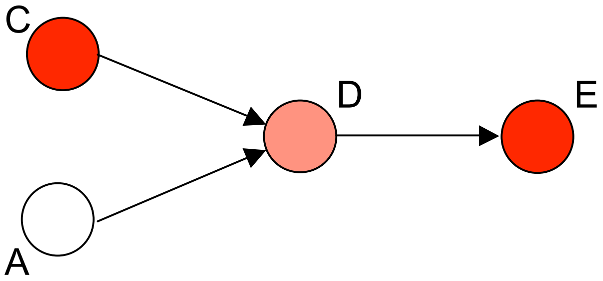


Figure 6

Remember that A's firing only makes a difference to whether D fires normally or feebly. Using causal language, we might put the point this way: event A does not cause event D, but does cause D to happen one way (as a normal firing) rather than another (as a feeble firing). There are also uses like the following: that C happens one way rather than another causes E. And finally: that C happens one way rather than another causes E to happen in one way rather than another. Here we see the default/deviant distinction intermixing with other contrasts that are more explicitly marked; in the last example, the default/deviant distinction is simply absent. Now, uses like these have led some authors (Lewis 1986c, Yablo 1992) to insist that we must distinguish, for example, *D's firing* from *D's firing normally*, in figure 5. These are *numerically different events*, so the story goes; in figure 6, only the former occurs, the latter being replaced by a new event, *D's firing feebly*.

Although I once found the arguments for such a multiplicity of perfectly coincident events persuasive, I now think they rest on a confusion about the kinds of causal explanations we can give, in answer to why-questions. Focus on some bit of the world, at some time. We can ask why that bit has such-and-such a state, at that time. Such questions are typically, and perhaps necessarily, contrastive: what we are really asking is why that bit has that state, rather than —, where the blank needs to be filled in somehow. There are two broadly different ways of filling it in. First choice: fill it in with the default state, for that bit of the world. Second choice: fill it in with some other state. I suggest that when we opt for the second choice, we almost always explicitly mark the intended contrast, somehow (sometimes with a "rather than" clause, sometimes with stress, etc.). If we do not mark the intended contrast by any explicit means, then the presumption is that this contrast is with the default state. At any rate, linguistics aside, there is clearly a useful distinction to be drawn between why-questions that contrast an actual state with a *default* state, and why-questions that contrast an actual state with an alternative *deviation*.

The very same distinction shows up on the end of *answers* to such why-questions, as well. Asked why some bit of the world had *that* state, rather than such-and-such an alternative state, we can reply that this other bit had *this* state, rather than such-and-such an alternative.[38] This alternative might be, on the one hand, the *default* state for the given bit of the world, or, on the other hand, some non-actual *deviation*. We thus have a four-fold division: two kinds of questions, two kinds of answers. I think our causal talk marks these divisions, in just the way we saw two paragraphs ago. And once they are clearly in view, there should be no temptation at all to think that our causal talk requires, for its proper understanding, the postulation of a teeming multitude of perfectly coincident events. To think *that* is to vastly overinflate the ontological significance of different ways of asking and answering why-questions. Seeing how

[38] Of course, we can give lots of other kinds of replies, as well.

the default/deviant distinction can interact with other distinctions helps bring this point into focus.

The third question recalls the methodological remarks in §1. There, I suggested that we treat intuitions about cases not as non-negotiable data, but rather as guides to where interesting and useful causal concepts can be found. The "reduction" account, if it succeeds, certainly articulates an interesting causal concept. But is it useful? For what, exactly? One possible answer: it is useful for the semanticist who wishes to provide systematic truth-conditions for English causal locutions. Maybe. Still, it would be nice to have another answer, another, more ambitious use to which we could put the concept the "reduction" account aims to articulate. There is, currently at least, something of a puzzle about what this use could be.

John Campbell has put the point rather nicely[39], in a way that shows that the puzzle doesn't really concern the "reduction" account or indeed any other account, but rather our causal intuitions themselves. It is perfectly understandable, Campbell notes, that we should conceive of the world in a way that makes that world easier for us to navigate: in reasoning about what to do, we need conceptual tools that will help us figure out how to *manipulate* bits of the world, so as to achieve our ends. It might be quite useful to know that if we do X, then Y is guaranteed to occur (or at least guaranteed to have a high probability of occurring), whereas if we don't do X then we can be sure of no such guarantee. (Maybe other forces are in motion that will guarantee Y's occurrence; still what we care about are the guarantees over future states of the world that *we* can effect.) The relevant notions of control, manipulation, guarantee and the like are surely causal, and perfectly good work in metaphysics could be done in giving precise articulations of them. But now consider Suzy First, again. Suzy's action guarantees that the bottle will shatter. So does Billy's. From the standpoint of control, manipulation, and guaranteeing future states of the world—from the standpoint, that is, of the causal notions that we can recognize as having a clear reason for being—they seem perfectly on a par. But they are *not* on a par—not, at least, as far as ordinary and quite unshakeable intuition is concerned. Why is that? What good does it do us to have such a firm tendency to want to distinguish the relation between Suzy's action and the bottle's shattering as being a profoundly different *causal* relation from that which holds between Billy's action and the shattering? That, I think, is an excellent question; it is the question that lurks behind the search for a good use to which to put the "reduction" account, or whatever account improves upon it.

[39] At a session devoted to recent philosophical work on causation, at the Pacific APA, March 2006.

Fourth, the role of the default/deviant distinction in our causal reasoning raises a fascinating question about the extent to which we can expect to be able to draw rich causal distinctions within any domain. Work with neuron diagrams, and you can distinguish, quite easily and clearly, between early preemption, late preemption, switching, short-circuits, threat-canceling, threat-canceling with backup, symmetric overdetermination, and no doubt many more varieties of causal structure. I suspect, though I do not yet know how to demonstrate, that it is the availability of a crystal clear, perfectly sharp default/deviant distinction that enables all of these distinctions to be drawn. More specifically, in many cases our conception of the causal structure of a situation informs us that the causal relationships between events are secured by the way that the *processes* or *mechanisms* those events are involved in interact.[40] I strongly suspect that this ability to discern a structure of interacting processes rests on a prior ability to distinguish default from deviant states of the relevant components.

Part of the reason is theoretical: the best account of what a "process" is (that I know of, anyway) says, roughly, that to distinguish which events constitute a "process" initiated by some events C_1, …, C_n that occur at time t, we must compare what would have happened, had the events C_1, …, C_n been the *only* events to occur at time t. (See Hall 2004a, 2004c.) Without a default/deviant distinction ready to hand, we can't make sense of this test. But part of the reason rests on considering certain odd examples, such as the multi-colored cellular automaton discussed in §6. That example was specifically designed to make it impossible to draw a default/deviant distinction. Now imagine some portion of the history of such a cellular automaton. Given some initial state, we can perfectly predict which states will follow it. We can even impose a kind of coarse causal gloss on this evolution: the state of each cell, at a time t, will be immediately caused by the states of it and the neighboring 8 cells, at time t-1; will be slightly less immediately caused by the states of it and the neighboring 24 cells, at time t-2; etc. But can we discern any *finer* causal structure than this? Well, yes, sometimes: for example, if cell A is red at time t, and would *not* have been red if neighboring cell B had had any other color but blue, at time t-1, then we can specifically nominate B's being blue at t-1 as a cause of A's being red at t. But we have to get lucky, as it were, to discern such structure—and I think that structure significantly *richer* than this is going to be impossible to discern. As a test, you might try to imagine a combination of rules for this cellular automaton, together with a sequence of states, such that it is crystal clear that the state of cell A at time t cancels a threat imposed by the state of cell B at t to the state of cell C at much later time t'. I predict

[40] Not necessarily in a simple way: e.g., it's not that we will judge C to be a cause of E iff there is a process connecting C to E. Cases of switching show that such a connection does not suffice for causation; cases of threat-canceling show that such connection is not necessary for causation.

that if you succeed, it will only be because you have settled on rules relative to which one cell state is a natural choice of default state (e.g., because it is the unique cell state such that if every cell has that state, then the rules guarantee that every cell will continue to have that state; see fn. 31, and see Maudlin 2004 for related discussion).

This suspicion, if correct, has relevance for real-world domains, notably *the mind*. People can have, at any given time, a rich set of beliefs, desires, intentions, etc. Let us grant that the having of such states can be thought of as the occurring of a large number of distinct mental *events*—not, presumably, because they occur in wholly distinct portions of the brain or soul, but perhaps because the relevant mental states can be varied independently of one another. (You could have this belief with this desire, or this belief with that other desire, etc.) Let us even grant that we can make good sense of counterfactual situations in which most of the mental events that actually occur in a given person at a given time are held fixed, while one of them is varied. (You have such-and-such beliefs, desires, intentions, etc.; consider what would have happened, had just this one belief been different in such-and-such a way….) I actually think we've probably granted way too much by this point, for reasons nicely articulated in Campbell (2006). Never mind. What would be *crazy* to grant—at least, without a great deal of supporting work from empirical psychology, none of which, to my knowledge, has been carried out—is that for any given mental event, there is a clear choice of default state—a clear and determinate conception of *what the mind would be doing instead, had that event not occurred*. Put simply, I suspect that the mind is much closer to the cellular automaton than it is to the typical neuron diagram. If so, that may make a profound difference to the questions about mental causation for which we can reasonably expect answers. Suzy goes to her favorite coffee shop. Why? Well, she reckoned she would find Billy there, and wanted to meet up with him. That was reason enough. But in addition, she craved espresso, and the coffee shop makes it just to her liking. That is *also* reason enough. "Fine," we might respond, "but which of these reasons was the *causally operative* one, on this occasion? Did the first preempt the second? Did the second preempt the first? Was this a case of symmetric overdetermination?"

I see no reason to think that these questions make any sense. But if they do, it will be in part because, surprisingly, investigation into human psychology reveals that there *is* a natural default/deviant distinction to be drawn. As opposed to pointless debates about the phony "exclusion problem", *this* seems to me a question about mental causation worth pursuing.

References

Arntzenius, Frank and Maudlin, Tim 2005: "Time Travel and Modern Physics," http://plato.stanford.edu/entries/time-travel-phys/.

Campbell, John 2006: "An Interventionist Approach to Causation in Psychology," in Alison Gopnik and Laura Schulz (eds.), *Causal Learning: Psychology, Philosophy and Computation*, Oxford: Oxford University Press, in press.

Clark, Peter and Hawley, Katherine eds. 2003: *Philosophy of Science Today*, Oxford: Oxford University Press.

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

Elga, Adam 2001: "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* 68: 3(Supplement): S313-S324.

Hall, Ned 2000: "Causation and the Price of Transitivity," *Journal of Philosophy* 97: 198-222.

Hall, Ned and Paul, L. A. 2003: "Causation and Preemption," in Clark and Hawley (eds.) *Philosophy of Science Today*, Oxford: Oxford University Press.

Hall, Ned 2004a: "The Intrinsic Character of Causation," in Dean Zimmerman (ed.), *Oxford Studies in Metaphysics, Volume 1*:255-300.

Hall, Ned 2004b: "Rescued from the Rubbish Bin: Lewis on Causation," *Philosophy of Science* 71: 1107-1114.

Hall, Ned 2004c: "Two concepts of causation," in Collins, Hall, and Paul 2004, *Causation and Counterfactuals*, chapter 9.

Halpern, Joseph Y., and Pearl, Judea 2005: "Causes and Explanations: A Structural-Model Approach. Part 1: Causes", *British Journal for the Philosophy of Science* 56: 843-887.

Hitchcock, Christopher 2001: "The Intransitivity of Causation Revealed in Equations and Graphs," *Journal of Philosophy* 98: 273 - 299.

Hitchcock, Christopher 2003: "Of Humean Bondage," British Journal for the Philosophy of Science 54: 1-25.

Kim, Jaegwon 1971: "Causes and Events: Mackie on Causation," *Journal of Philosophy* 68: 426-41.

Lewis, David 1973a: "Causation," *Journal of Philosophy* 70:556-567.

Lewis, David 1973b: *Counterfactuals*, Cambridge: Harvard University Press.

Lewis, David 1979: "Counterfactual Dependence and Time's Arrow," *Noûs* 13: 455-476.

Lewis, David 1986a: *Philosophical Papers, Volume II*. New York: Oxford University Press.

Lewis, David 1986b: Postscripts to "Causation," in Lewis 1986a: 172-213.

Lewis, David 1986c: "Events," in Lewis 1986a: 241-269.

Lewis, David 2004: “Causation As Influence,” in Collins, Hall, and Paul 2004, *Causation and Counterfactuals*, chapter 3.

Maudlin, Tim 2003: “A Modest Proposal Concerning Laws, Counterfactuals, and Explanation,” unpublished ms.

Maudlin, Tim 2004: “Causation, Counterfactuals, and the Third Factor,” in Collins, Hall, and Paul 2004, *Causation and Counterfactuals*, chapter 18.

McDermott, Michael, 1995: “Redundant Causation,” *British Journal for the Philosophy of Science* 46: 523-544.

Pearl, Judea 2000: *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

Ramachandran, Murali 1997: “A Counterfactual Analysis of Causation,” *Mind* 106:263-277.

Yablo, Stephen 1992: “Cause and Essence,” *Synthese* 93: 403-449.

Yablo, Stephen 2004: “Advertisement for a Sketch of an Outline of a Proto-Theory of Causation,” in Collins, Hall, and Paul 2004, *Causation and Counterfactuals*, chapter 5.

Zimmerman, Dean ed. 2004: *Oxford Studies in Metaphysics, Volume 1*, Oxford:Clarendon Press.