Philos Stud (2007) 132:109–136
DOI 10.1007/s11098-006-9057-9

ORIGINAL PAPER

Structural equations and causation

N. Hall

Received: 23 October 2006 / Accepted: 25 October 2006 /
Published online: 13 January 2007

Abstract Structural equations have become increasingly popular in recent years as tools for understanding causation. But standard structural equations approaches to causation face deep problems. The most philosophically interesting of these consists in their failure to incorporate a distinction between default states of an object or system, and deviations therefrom. Exploring this problem, and how to fix it, helps to illuminate the central role this distinction plays in our causal thinking.

Keywords Causation · Counterfactuals · Causal models · Structural equations · Defaults · Deviants

1 Introduction

Among philosophers and scientists interested in causation, the idea has gained great currency 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 appropriate *variables*, together with *structural equations* that capture the relations of dependence among them. The key advantage—what, in the eyes of at least some authors, makes these models 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 elusive.[2]

[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 and Pearl (2005).

N. Hall (✉)
Department of Philosophy, Harvard University, Cambridge, MA, USA
e-mail: ehall@fas.harvard.edu

This is 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 fascinating lessons about causation by showing why.

Some reasons for pessimism leap out once you focus on two obvious questions:

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

Regrettably, for reasons of space we'll have to pass over these questions, sticking to examples straightforward enough that they do not arise so urgently. (But see Hall 2006—the extended version of this paper from which the present version is extracted—for extensive discussion. Hereafter, "the extended version".) I will simply note that 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.

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 advertised when applied to canonical preemption examples, etc. (Again, see the extended version.) 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.[3] (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 easy to produce a vastly improved structural equations account. (Better: an improved account that could, if one liked, be presented within a structural equations framework.) So debunking some of the hype surrounding the structural equations approach to causation will at least point 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.

2 Some simple examples

Let's start with some examples that work very well to give the flavor of structural equations approaches. 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.

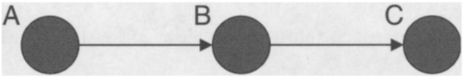


Fig. 1

[3] 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.

Here, 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 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. Throughout, I'll use capital letters interchangeably to refer to neurons and to events of their firing.

There is no mystery about what causes what in a situation like that depicted in Fig. 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 clear tests for any analysis of causation. But they 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 Fig. 1, because it's too simple. So consider Fig. 2:

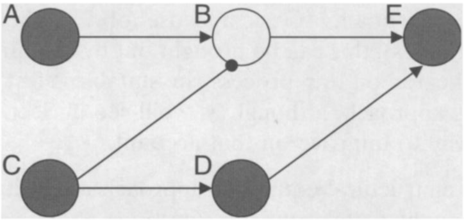


Fig. 2

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. Many real-world situations have this simple preemptive structure.

As is well-known, examples like Fig. 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 Fig. 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 Fig. 2. Here are the main options:

- Even though E does not depend on C, it *does* depend on D, and D on C;[4] 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 1973a, b.)
- 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.

[4] 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 the extended version, Sect. 4.

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 and 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 connecting 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 Sect. 4, 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 Figs. 1 and 2. In doing so, we really ought to take up the questions about variables and structural equations raised above. But we can get away with ignoring them, thanks to the highly sanitized nature of neuron diagrams. For example, in modeling Fig. 1 it is more or less obvious that we should choose three binary variables:

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

There is nothing special about the numbers 0 and 1; they are mere labels.

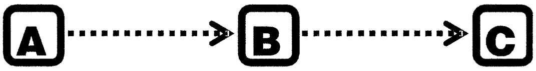
Next, it is more or less obvious how to write down structural equations that capture the relations of immediate dependence between these variables:

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

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

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 **C** is to be "set" to the same value as **B**). In general, for any variable **X** in any given model, the structural equations for that model will distinguish those other variables that **X** depends on (either immediately or mediately) from those it doesn't: **X** will depend on **Y** iff there is a sequence of variables **Y**, $\mathbf{Z_1}$, $\mathbf{Z_2}, \ldots, \mathbf{Z_n}$, **X** such that **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 $\mathbf{Z_2}, \ldots, \mathbf{Z_n}$ appears on the right-hand side of the structural equation for **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, **A** is the sole exogenous variable.)[5]

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



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 Fig. 1, this directed graph should obviously not be *confused* with Fig. 1. (For example, only Fig. 1 contains a depiction of what *actually happens*.)

Virtually every structural equations account of causation will say the same thing about why A, in Fig. 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:

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

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

$$\mathbf{A} = \mathbf{0}$$

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

$$\textbf{if } \mathbf{A} = \mathbf{0}, \textbf{ then } \mathbf{C} = \mathbf{0}$$

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

$$\textbf{If } \mathbf{X} = \mathbf{v}, \textbf{ then } \mathbf{P}$$

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 on **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

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

results then determines the truth of **P**, and so the truth of the conditional. Since **A** does not depend on **C**, we have

$$\textbf{if } \mathbf{C} = \mathbf{0}, \textbf{ then } \mathbf{A} = \mathbf{1}$$

Hence C does not come out a cause of A.

Some comments.

First, in the more general treatment (see Halpern and Pearl, 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 "$\mathbf{X} = \mathbf{v}$". I'll 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 need suitable translations of claims to the effect that some event occurs into the language of the model. Neuron diagrams are easy; but you should be warned that 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 the system of structural equations for a given model might contain *loops*, so that distinct variables **X** and **Y** depend on *each other*. If so, it can happen that, for a particular choice $\mathbf{X} = \mathbf{v}$, there is no way to update the values of the variables that depend on **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.[6] So I'll assume, henceforth, that our causal models behave themselves, and never feature such loops.

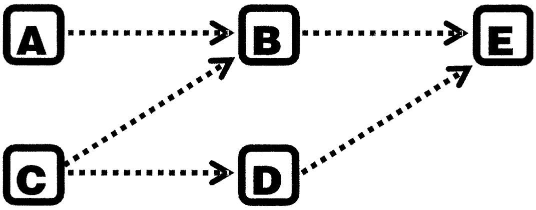
Third, this account of conditionals will remind you of the requirement, standard in counterfactual analyses of causation, that the counterfactuals used in the analysis be given a *non-backtracking* reading. You might therefore suspect the need for a story—perhaps involving Lewis's (1979) "miracles"—that 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 come as no surprise that answering that question will revive the issue of "miracles" (see the extended version, Sect. 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 survive scrutiny. The main reason is that decent truth-conditions for the structural equations 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 assume that's not a problem; see the extended version, Sect. 4, for a vindication of this assumption.

[6] 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 and Maudlin (2005) for relevant details.

Finally, one might wonder what the big deal is, if the example of Fig. 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 Fig. 2.

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



And here are the structural equations:

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

Finally, the actual values are these:

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

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

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

is false, the following conditional is *true*:

$$\textbf{if } (\mathbf{C} = \mathbf{0} \;\&\; \mathbf{B} = \mathbf{0}), \textbf{ then } \mathbf{E} = \mathbf{0}$$

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 $\mathbf{X_1,\ldots,X_n}$ with actual values $\mathbf{v_1,\ldots,v_n}$, the conditional

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

is true. It's easy to check that this account not only delivers the verdict that C in Fig. 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.

The account just sketched displays one of many options for using causal models in a theory of causation. We can usefully contrast a second option, simplifying the approach taken in Halpern and Pearl (2005). We first 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 $\mathbf{X_1,\ldots,X_n}$ and (not necessarily actual) values $\mathbf{v_1,\ldots,v_n}$, the conditional

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

is true. Of course, that's *too* liberal: for example, it counts A in Fig. 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:

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

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.[7]

Now A in Fig. 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 **C = 0**. And while

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

is true,

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

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 Fig. 3:

[7] 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.

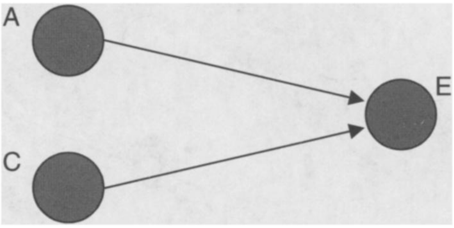


Fig. 3

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.

So far, the enthusiasm toward structural equations approaches might seem justified, given their novel, interesting, and effective means for treating certain preemption cases. But the examples that follow tell a different story.

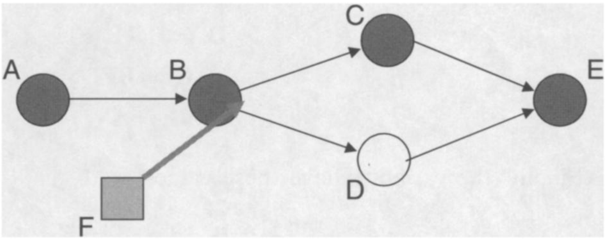
3 Trouble cases

Let's look at three more examples.[8]

3.1 Switches

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 Fig. 4, then the stimulatory signal gets sent to C; if it doesn't, the signal gets sent to D:

Fig. 4



[8] The extended version looks at structural equations treatments of late preemption, as well—too long a discussion to include here. But I cannot resist observing that there is an astonishing gap—a chasm, a Grand Canyon—between the claims that partisans of structural equations make on behalf of these treatments, on one hand, and the fallacy-ridden reality, on the other.

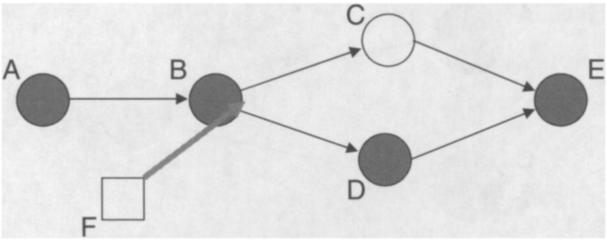
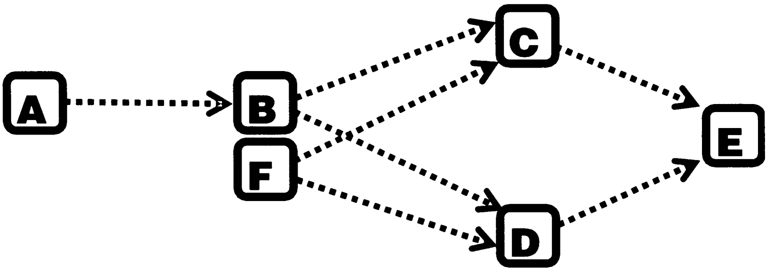


Fig. 5

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 Fig. 4, is not a cause of E—notwithstanding that it *is* a cause of C, and C of E.[9] 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:



Here are the equations:

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

Finally, the variables have these values:

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

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

[9] 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.

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

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.[10]

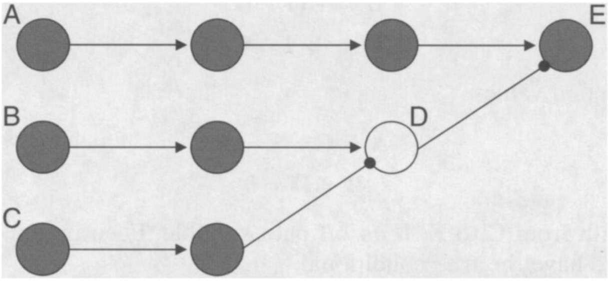


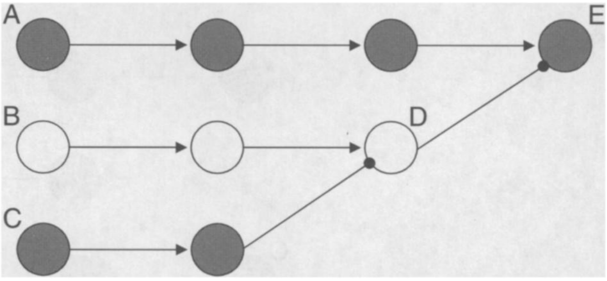
Fig. 6

3.2 Non-existent threats

Figure 6 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.[11] 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 Fig. 7:

Fig. 7



[10] Halpern and Pearl (2005) think otherwise, offering a rather tortured defense of the claim that *other* acceptable models for switches will yield the result that the switching event is not (relative to those models) a cause of the target effect. See the extended version for discussion.

[11] Thus I am distancing myself somewhat from the view expressed in Hall (2004c), though largely to avoid needless complication.

In Fig. 6, 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 Fig. 7, there is no such threat. It would therefore be absurd to count C a cause of E. 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 a silly conflation of *causing* with *safeguarding*.

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 Fig. 7, with these equations:

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

We have the actual values

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

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

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

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

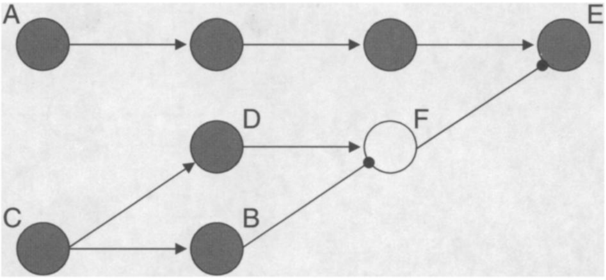
$$\textbf{if } (\mathbf{C} = \mathbf{1}\ \&\ \mathbf{B} = \mathbf{1}),\ \textbf{then } (\mathbf{C} = \mathbf{1}\ \&\ \mathbf{D} = \mathbf{0}\ \&\ \mathbf{E} = \mathbf{1})$$

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

3.3 Short-circuits

The H-account, at least, does not fall into the trap set by Fig. 7. But it (hence the HP-account as well) does fall into a closely related trap:

Fig. 8



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.

C is not a cause of E.[12] 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 Fig. 8 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, together with these 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

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

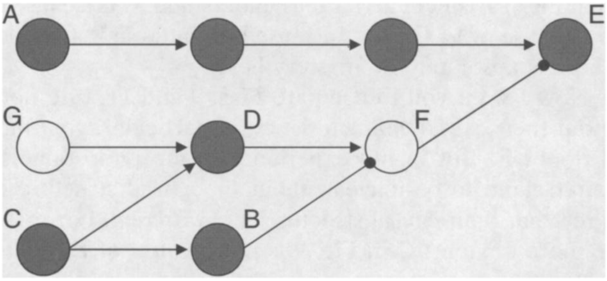
We ignore the third equation. Having set **C** to 0 and **D** to 1, **B** = 0 by the fourth equation. Then **F** = 1 by the second equation, and thus **E** = 0 by the first. So the conditional is true. So both the H-account and HP-account classify C as a cause of E—for *exactly the same reason* that they count C, in Fig. 2, a cause of E. (Whence we now have good reason to doubt that they got that case right *for the right reasons*.)

Time to unearth the deep errors committed by structural equations accounts that lie behind the trouble exhibited in this section.

3.4 The default/deviant distinction

Let us examine two different cases, with markedly different causal structures. The first is a variant on the 'short-circuit' of Sect. 2.3:

Fig. 9

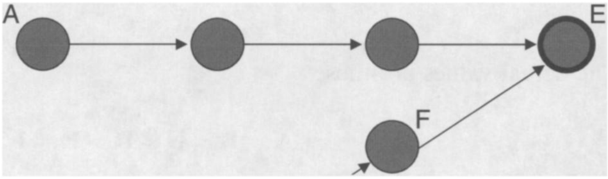


[12] For a rare—and strained—disagreement, see Lewis (2004).

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*.

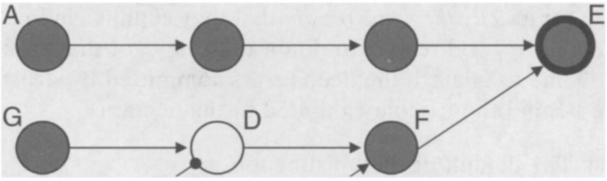
Maybe you don't agree. Never mind. All that really matters, for present purposes, is that we see clearly that, whatever causal structure we might wish to impute to the events in Fig. 9, it should be a *different* causal structure from that exhibited by the next case. We will build up to that case in stages:

Fig. 10



Neuron E in Fig. 10 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:

Fig. 11



Look at the little network G–D–C–B–F. You've seen it before, in Fig. 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 Fig. 11, although more complicated than Fig. 10, 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 Figs. 9 and 11. I do not ask that you agree with me about their causal characteristics; in particular, you might find Fig. 9 too confusing. (I doubt it. But in my experience, some philosophers who really ought to know better claim to be unclear about the causal structure of Fig. 9.) I *do* ask that you agree that their causal structures are *different*. For myself, one difference could not be more obvious: C in Fig. 9 is not a cause of E; C in Fig. 11 *is* a cause of E. But again, it's enough that you recognize that *some* difference exists. 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 Fig. 9, our model will include the obvious seven binary variables. Here are their equations:

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

The model for Fig. 11 will likewise include seven binary variables, this time with these equations:

$$\mathbf{E} \Leftarrow \mathbf{AF}$$
$$\mathbf{F} \Leftarrow \mathbf{D}+\mathbf{B}-\mathbf{DB}$$
$$\mathbf{D} \Leftarrow \mathbf{G}(\mathbf{1}-\mathbf{C})$$
$$\mathbf{B} \Leftarrow \mathbf{C}$$

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 Fig. 9 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:

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

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

Accordingly, let's rewrite the equations in the model for Fig. 11, using new variables in place of **D, F**, and **G**:

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

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

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

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

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

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

That the two models are in fact the same is now obvious. Finally, observe that in Fig. 9, the actual values are these:

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

In Fig. 11, the actual values are these:

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

Suppose an account of causation tries to render a verdict about what causes what, in Fig. 9, making use *just* of the structural equations for Fig. 9, plus the actual values of the variables. Suppose that account tries to do the same, for Fig. 11. Then the isomorphism between the models establishes—*conclusively*—that the account will call C in Fig. 9 a cause of E iff it likewise calls C in Fig. 11 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 Fig. 9, the state neuron F is in, at the relevant time, is a *non-firing* state; in Fig. 11, 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. More generally, 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 acting on 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. Obviously, if nothing else exists, then nothing acts on the particle. So we have our test environment, non-causally characterized, and can use it to define default behavior for a Newtonian particle as 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 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* (at least, on *one* legitimate way to draw the default/deviant distinction). 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!).

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 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, one important role for the default/deviant distinction derives from 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 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? It seems that that question should direct our attention to a situation whose character, as regards the switch's behavior, is highly indeterminate. But it doesn't: 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*.

Contrast 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 assignment of a default state is impossible. As an artificial but vivid example, consider a cellular automaton in which each cell can have, at each moment, one of four 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.[13] 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 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. We could 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

[13] 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.

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. That's a phenomenon well worth further exploration. Here, though, we'll stick to easier cases, where the 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 Sect. 2.

4 An improved account

The account I will offer makes use of the following 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'.

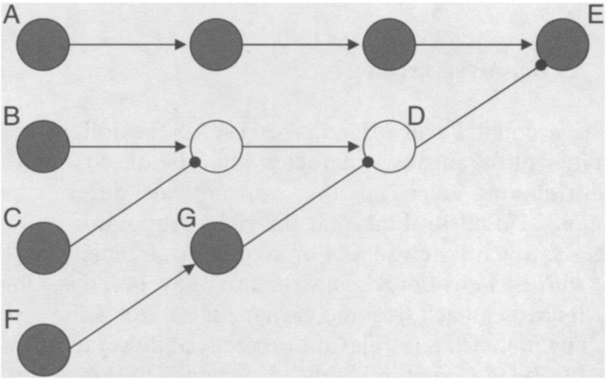
I used to think that Intrinsicness provided the key to one paradigmatic kind of causation—what I called "production"—in which the cause brings about its effect by way of a connecting process. Production, I thought, should be contrasted with *dependence*, a more minimal kind of causation in which the 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, typical preemption cases would exhibit production without dependence, cases of threat-canceling dependence without production.

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 *can* conflict: think of "child", or "mother". 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 Sect. 2.1. 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 result that even in the original case, the engineer's action is a cause of the arrival. Generalizing, an account of production that rests on Intrinsicness must call switches *producers* of the relevant effect, and so in one central sense *causes*.

The second problem arises from variants on threat-cancelling, in which a *backup* threat-canceller is present, but remains idle. Figure 12 illustrates:

Fig. 12



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 count as a producer of E.[14] 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.

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; 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 dependence, 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

[14] 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.

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 too easy to pass. Our test isn't. I thus 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.[15] 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.[16] 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 sit-

[15] *Exactly* the same state? No. See the extended version for discussion of this qualification, and the reasons for it.

[16] 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.

uation, 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

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

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 Fig. 2:

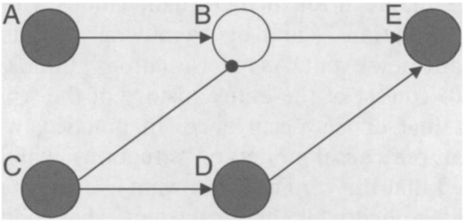


Figure 2

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

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

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.

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

$$\textbf{if } \mathbf{A} = \mathbf{0}, \textbf{ then } \mathbf{E} = \mathbf{0}$$

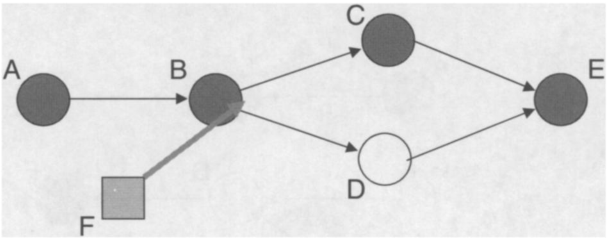
is false. The only variant in which it is true is the one in which **A = 1** and **C = 0**. But this is not a *reduction* of the actual situation: for **B** has the value 1, which is neither its default value nor its actual value.

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 Fig. 2, by omitting the variables **B** and **D**. Our one equation would then be

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

According to this model, both A and C are causes of E. No surprise: this model effectively (mis)treats Fig. 2 as a case of symmetric overdetermination. Now, we already knew that this was a bad model for Fig. 2. But now we can say 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:



Figuer 4

Here are the equations:

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

The variables have these actual values:

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

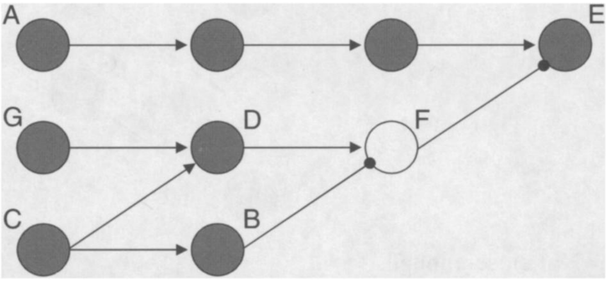
A and **F** are the sole exogenous variables. To find a reduction in which E depends on F, we must of course let **F = 1**. Then the only variation we can construct is the one in which **A = 0** and **F = 1**. But then **E = 0**. So the model does not even yield a *variation* in which E depends on F, let alone a reduction in which it does so. So F is not a cause of E.

Next, non-existent threats. Here, a glance back at Fig. 7 will confirm that there is no reduction in which E depends on C; so no event will count as a cause simply because it offers safeguards against a non-existent threat.

Next, short-circuits. As with switches, there is no variant in which E depends on C, hence no reduction in which it does so.

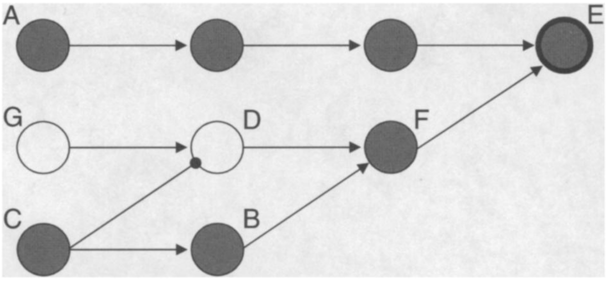
Next, let us compare Figs. 9 and 11; we won't stop to reproduce the causal models. Figure 9 has one variant in which E depends on C:

Fig. 13



But this variant is not a *reduction*, since G is in neither its default state nor its actual state. Figure 11, by contrast, *does* have a reduction in which E depends on C:

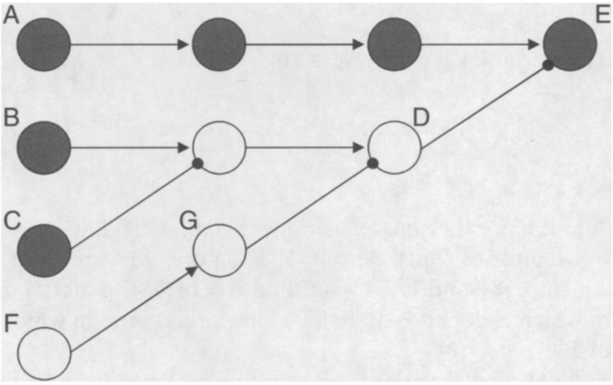
Fig. 14



The account thus neatly secures the obvious contrast between the causal structures of Figs. 9 and 11.

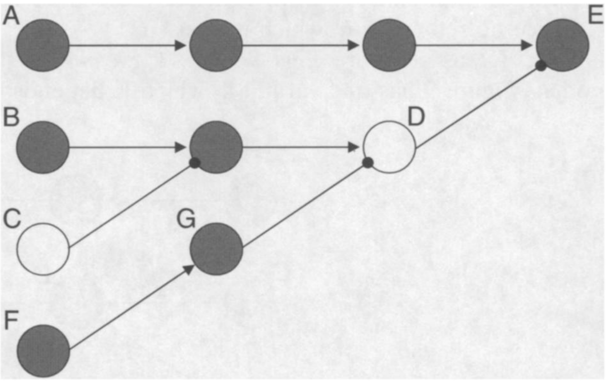
Next, threat-canceling with backup. Again, the contrast is easy to see. In Fig. 15, we have a reduction of the situation depicted in Fig. 12, and E depends on C:

Fig. 15



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

Fig. 16



Not close enough.

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—as we've just seen—possible that improve dramatically on the 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 early 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 further questions about this distinction that strike me as most worth investigating.

5 Some larger questions

First, what makes the distinction tick? In Sect. 3, I offered some sketchy remark on this topic, but a proper theory would be welcome. I suggest that 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 Fig. 17:

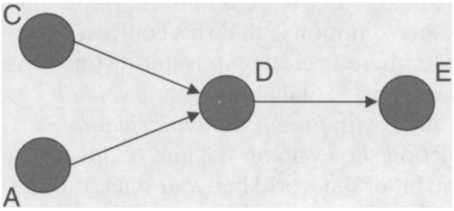


Fig. 17

The connection between A and D in Fig. 17 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 Fig. 17, or feebly, as in Fig. 18:

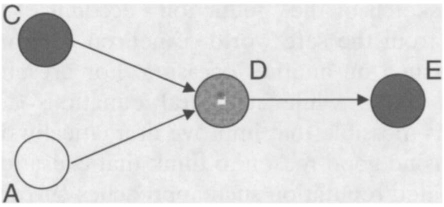


Fig. 18

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.

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 (not illustrated by Figs. 17 and 18): 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 Fig. 17. These are *numerically different events*, so the story goes; in Fig. 18, 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. 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 the default/deviant distinction can interact with other distinctions helps bring this point into focus.

Third, 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.[17] 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.

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 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—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*. 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 was *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 be certain that these questions make any sense. But if they do, it will be in part because, surprisingly, investigation into human psychology reveals

[17] Not in a simple way: 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 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, F., & Maudlin, T. (2005). Time travel and modern physics. http://plato.stanford.edu/entries/time-travel-phys/

Campbell, J. (2006). An interventionist approach to causation in psychology. In: A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy and computation*. Oxford: Oxford University Press, in press.

Clark, P., & Hawley, K. (Eds.). (2003). *Philosophy of science today*. Oxford: Oxford University Press.

Collins, J., Hall, N., & Paul, L. A. (Eds.). (2004). *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science, 68*(3)(Supplement), S313–S324.

Hall, N. (2000). Causation and the price of transitivity. *Journal of Philosophy, 97*, 198–222

Hall, N., & Paul, L. A. (2003). Causation and preemption. In: P. Clark & K. Hawley (Eds.), *Philosophy of science today*. Oxford: Oxford University Press.

Hall, N. (2004a). The intrinsic character of causation. In: D. Zimmerman (Ed.), Oxford Studies in Metaphysics, Volume 1:255–300.

Hall, N. (2004b). Rescued from the rubbish bin: Lewis on causation. *Philosophy of Science, 71*, 1107–1114.

Hall, N. (2004c). Two concepts of causation. In: J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals*, chapter 9.

Hall, N. (2006). Structural equations and causation (extended version) ms.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part 1: Causes. *British Journal for the Philosophy of Science, 56*, 843–887.

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

Hitchcock, C. (2003). Of humean bondage. *British Journal for the Philosophy of Science, 54*, 1–25.

Kim, J. (1971). Causes and events: Mackie on causation. *Journal of Philosophy, 68*, 426–441.

Lewis, D. (1973a). Causation. *Journal of Philosophy, 70*, 556–567.

Lewis, D. (1973b). *Counterfactuals*. Cambridge: Harvard University Press.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*, 455–476.

Lewis, D. (1986a). *Philosophical papers* (Vol. II). New York: Oxford University Press.

Lewis, D. (1986b). Postscripts to "Causation." In: Lewis 1986a: 172–213.

Lewis, D. (1986c). Events. In: Lewis 1986a: 241–269.

Lewis, D. (2004). Causation as influence. In: J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (chapter 3).

Maudlin, T. (2003). A modest proposal concerning laws, counterfactuals, and explanation. unpublished ms.

Maudlin, T. (2004). Causation, counterfactuals, and the third factor. In: J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals*, (chapter 18).

McDermott, M. (1995). Redundant causation. *British Journal for the Philosophy of Science, 46*, 523–544.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Ramachandran, M. (1997). A counterfactual analysis of causation. *Mind, 106*, 263–277.

Yablo, S. (1992). Cause and essence. *Synthese, 93*, 403–449.

Yablo, S. (2004). Advertisement for a sketch of an outline of a proto-theory of causation. In: J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (chapter 5).

Zimmerman, D. (Ed.) (2004). *Oxford studies in metaphysics* (Vol. 1). Oxford: Clarendon Press.