Levels of Explanation in Woodward’s Theory of Causal Explanation

Bradford Skow

Abstract

Woodward’s theory says that to causally explain why E occurred one must exhibit, or imply, some counterfactual concerning E. I will argue that causal explanations do not need to do this. But Woodward’s theory is close to the truth. There are different levels of explanation. At the first level there is explaining why E occurred; at the second, explaining why C explains why E occurred. Exhibiting the truth of a counterfactual, while not necessary at the first level, is necessary at the second.

1 Introduction

James Woodward’s book Making Things Happen (2003) develops both a theory of causation and, based on it, a theory of causal explanation. My interest here is in the theory of causal explanation. Since the book’s publication over ten years ago this theory has become one of the most influential theories of causal explanation available. This clearly makes it worthy of critical attention. Although the theory is influential, I think it is false.

Most critical discussions of Woodward’s theory that I am aware of do not challenge his basic core claims about causal explanation. They focus on Woodward’s extensions of, or applications of, those basic claims. For example, Weslake (2010) and Franklin-Hall (forthcoming),

1All references to Woodward’s work are to this book unless otherwise noted.
2There are fewer discussions of Woodward’s theory of causal explanation than one might
in different ways, dispute Woodward’s claim that his theory is compatible with the fact that the explanations found in the special sciences are in some sense deeper than, or better than, corresponding explanations in the language of fundamental physics. But I think Woodward’s theory runs into trouble before we get to sophisticated examples that compare, say, thermodynamic to microphysical explanations. I think it says the wrong things about even very simple causal explanations. In brief: Woodward’s theory says that to causally explain why E occurred one must exhibit, or imply, some counterfactual concerning E; but causal explanations do not need to do this.

But Woodward’s theory is close to the truth. There are different levels of explanation. At the first level there is explaining why E occurred. It might be that C explains why E occurred. We might then immediately ask why this is so. This takes us to the second level: explaining why C explains why E occurred. I will argue that exhibiting the truth of a counterfactual, while not necessary at the first level, is necessary at the second.

2 A Simple Argument Against Woodward’s Theory

My argument is simple. I will give it here in outline, and defend it at length in section 5. Suppose lightning strikes in a field, causing the field to catch on fire. Then this is certainly a causal explanation of the fire:

(1) The field caught on fire because it was struck by lightning.

But, I will argue, Woodward’s theory does not count this as a causal explanation. Therefore, Woodward’s theory of causal explanation is false.

A common reaction to my argument is that it cannot be right. It cannot be that easy to show that Woodward’s theory is false. After all, doesn’t Woodward spend several pages of *Making Things Happen* discussing causal explanations like (1)? If he paid that much attention expect. Google Scholar’s list of citations of Woodward’s book contains very many entries on his theory of causation and very few on his theory of causal explanation.
to them, how could he fail to make his theory accommodate them? Let me try to convince that he
did fail. I will start with a statement of Woodward’s theory, and then look at what he says about
examples like (1).

3 A Statement of Woodward’s Theory

I find it helpful to contrast Woodward’s theory with a very naive theory of causal explanation.
This naive theory just says that to causally explain why some event E occurred it is necessary
and sufficient to describe one or more causes of E. (It is worth noting that example (1) is clearly
a causal explanation, according to the naive theory.) Woodward’s theory is more sophisticated.
But why? What is wrong with the naive theory?

We can find an answer by looking at Woodward’s theory of causation. It is a version of a
“structural equations” theory of causation.\textsuperscript{3} To apply the theory we model the causal facts of a
situation using (i) a set of variables and (ii) a set of equations. Let me start with variables. Each
variable represents some feature of the situation being modeled. It is easiest to introduce the
notion using a simple example. Modifying an example of Woodward’s, suppose that a lightbulb
is connected to a dial by a wire.\textsuperscript{4} The dial determines the current through the wire, and the
brightness of the bulb is in turn determined by the current. In fact the bulb is quite sensitive so
that any increase in the current causes an increase in the brightness. We could represent this
situation using two variables, D for the position of the dial, and B for the brightness of the bulb.\textsuperscript{5}

\textsuperscript{3}Theories of this kind draw heavily on (Pearl 2000) and (Spirtes et. al. 2000). Woodward
presents his theory in Chapters 2 and 3 of Making Things Happen. Woodward has had a lot
to say about causation in later publications (for example, Woodward 2006), but the theory in
the book is enough for our purposes. In fact the book discusses two different theories of token
causation; I will focus on the one labeled (AC) (2003: 77). Woodward does not claim that this
theory is original with him. He credits (Hitchcock 2001), (Pearl 2000), and (Halpern and Pearl
2000). By calling it Woodward’s theory I mean only to claim that it is one that Woodward
defends, not one he invented.

\textsuperscript{4}Woodward’s lightbulb example appears on page 66.

\textsuperscript{5}We could also include a variable for the strength of the current, but nothing would be gained
in this context from doing so.
The variable D can take on any real number between 0 and 180 as a value. This corresponds to the fact that the dial can be turned up to 180 degrees from its starting point, the “off” position. The variable B can take on any real number between 0 and 1 as a value, where 0 represents the bulb being off, and 1 represents the bulb at its maximal brightness.

What about equations? In any causal model there are two sets of variables: the endogenous variables, and the exogenous variables. Intuitively, the endogenous variables are effects of other variables in the model, and the exogenous ones are not. For each endogenous variable V there is an equation \( V = \text{F}(X_1, X_2, ..., X_n) \) determining its value as a function of the values of other variables Xi. In our example model there is only one endogenous variable, B, so we need only one equation. (The value of D influences the value of B, but there is no variable in the model that influences D.) If we suppose that the brightness of the bulb increases linearly as the current increases, then the equation for B is simply \( B = \frac{D}{180} \). Perhaps in the actual world the dial is at 0 degrees and the bulb is off. Then D and B both have the value 0.

This is a very simple example. The theory’s interest as a theory of causation comes in its application to more complicated models. But this example is enough to illustrate its relevance to the theory of causal explanation.

It is important to note that not just any collection of variables and equations counts as an eligible causal model. A set of variables and equations determines a directed graph, where each node in the graph corresponds to one of the variables, and there is an arrow from variable V to variable W iff V appears on the right-hand side of the equation that assigns values to W. (The graph for the simple example is just \( D \rightarrow B \).) A set of variables and equations is an eligible causal model iff it is acyclic: there is no path through the graph from any node to itself. (A path through a directed graph is an ordered sequence of nodes \((X_1, X_2, ..., X_n)\) such that there is an arrow in the graph from each member to the next.)

Woodward’s theory of (token) causation gives necessary and sufficient conditions for \( V = v \) to be a cause of \( W = w \), given an eligible causal model and a specification of the values of its...
variables. In brief, and omitting some details, his theory says: $V=v$ is a cause of $W=w$ iff there is some path from $V$ to $W$, and some possible value $v'$ of $V$ distinct from $v$, such that, “holding fixed” the variables not on that path at their actual values, had $V=v'$, $W$ would have taken some value other than $w$.

The theory’s application to our simple example is straightforward. $D=0$ is obviously a cause of $B=0$, since the sequence $(D,B)$ is a path from $D$ to $B$, and there are no off-path variables, and had $D$ taken on any value other than 0, the value of $B$ would not have been 0.

Again, this theory of causation gains its interest from its application to more complicated causal models. But even with this simple example we can see why Woodward might have been dissatisfied with the naive theory of causal explanation. The naive theory says that we can explain why the bulb is off by saying that the dial is off (in variable-speak, we can explain why $B=0$ by saying that $D=0$). If we exclude from consideration all other causes of the bulb’s being off, then the naive theory says that there is nothing more to be said about why the bulb is off. Woodward maintains that this is wrong. There is further causal-explanatory information to be had. For, as we have seen, according to Woodward’s theory of causation the claim

$$(2) \ D=0 \text{ is a cause of } B=0.$$ 

is equivalent to

---

6To be complete the theory needs to say what it takes for a causal model to correct represent some real-world state of affairs; see (Hall 2007) for discussion.

7Let me repeat again that this is only one of two theories of token causation that Woodward presents in his book. Woodward also discusses several notions of type-level causation. Those notions will not play a role in what follows.

A more complete statement of this theory of token causation requires the notion of an intervention. Woodward requires that the off-path variables be held at their actual values as a result of interventions, and that $V$ be set at value $v'$ as a result of an intervention. The precise definition of “intervention” is complex, and the kinds of problems that using the language of interventions is meant to solve will not arise in this paper.

8Of course my description of the dial-lightbulb scenario built-in that the dial’s being off caused the lightbulb to be off. But Woodward’s theory is not a reductive theory of causation; it does not require that the situation one is modeling be described in non-causal terms, or that the justification of one causal model as correct be done in non-causal terms.
(3) There is a path from D to B, and a number \( r \) with \( 0 < r \leq 180 \), and a number \( s \) greater than 0, such that, had D taken value \( r \), and the off-path variables had their actual values, B would have taken value \( s \).

(3) is a (triple) existential generalization. It says that at least one counterfactual in a certain class of counterfactuals is true. But it does not say which one. Clearly one could say more. One could say which of those counterfactuals are true, and which are false. One could say, for example, that had D taken the value 90, then B would have taken the value 1/2 (there are no off-path variables to worry about in this example). Woodward wants to count this more specific counterfactual information as causal-explanatory information. It is part of the answer to the question why \( B=0 \), a part of the answer that goes beyond the proposition that \( D=0 \) is a cause of \( B=0 \).

I have been trying to get the idea behind Woodward’s theory of causal explanation across in my own words. What I have said fits with Woodward’s informal characterizations of his theory. Here are a few of them:

(4) [causal] explanation is a matter of exhibiting systematic patterns of counterfactual dependence (191).

(5) the underlying or unifying idea in the notion of causal explanation is the idea that an explanation must answer a what-if-things-had-been-different question, or exhibit information about a pattern of dependency (201).

(6) The theory of causal explanation I have been sketching is thus one that ties explanatory import very closely to the provision of certain kinds of counterfactual information: it might fairly be described as a counterfactual theory of causal explanation (196).

All of this fits with what I have been saying about the example. Saying something about how the value of B depends on the value of D, either by making a specific claim about what value B would have had if D had had some other given value, or just by saying that B would have taken
some different value or other if D had, counts as providing causal-explanatory information, according to (4), (5), and (6).⁹

4 Woodward on Singular Causal Explanation

Go back to the example I started with:

(1) The field caught on fire because it was struck by lightning.

This is obviously a causal explanation. Is it a causal explanation, according to Woodward’s theory? He says quite a bit about examples like (1). At the beginning of section 5.8, on singular causal explanation, Woodward writes that “[t]oken-causal claims imply various counterfactuals, and it is in virtue of conveying this counterfactual information that we should think of them as explanatory” (210). He then discusses several examples very similar to (1). He asks, “what can we say about [their] counterfactual import..., or about what they tell us by way of answers to what-if-things-had-been-different questions” (211)? He answers that, if we assume that there is no causal overdetermination (for example, no preempted backup causes), then “a singular causal claim (or explanation) of form c caused e implies the following counterfactual: if c had not occurred, then e would not have occurred” (211). In our example, if we assume that there were no arsonists around waiting to light the field on fire if lightning did not strike, then, Woodward would say, “the lightning caused the fire” implies that, had the lightning not struck, the field would not have caught on fire. Woodward then maintains that “[w]e can think of this counterfactual information just as it stands as conveying information about the answer to what-if-things-had-been-different questions and as (at least minimally) explanatory for just this reason” (211).

⁹Well, not quite. (4) seems to require that a causal explanation exhibit a systematic pattern of dependence. I take it from what Woodward says about singular causal explanations that he does not in fact deem this necessary. (I discuss Woodward’s claims about singular causal explanations below.)
Of course (1) could be true even if there were an arsonist waiting in his car to light the field on fire if the lightning storm doesn’t do it for him. If there were such an arsonist then (1) would be true, but would not imply the counterfactual “had the lightning not struck, the field would not have caught on fire.” Woodward is aware of this. He would say that in this new scenario, (1) counts as a causal explanation in virtue of implying the truth of a different counterfactual, namely “Had lightning not struck, and the arsonist not acted, the field would not have caught on fire.” (See the discussion on 219.)

So far Woodward has been taking the examples case by case. In one situation (1) is a causal explanation by virtue of implying one counterfactual, in another by conveying a different counterfactual; what exactly is the theory here? It is easy to see what is going on here at at least one level of generality:

(7) A body of fact constitutes a causal explanation of some event E if and only if that body of fact implies the truth of at least one counterfactual of a particular kind K.

What is kind K? A counterfactual C that mentions the fact that V=v in its antecedent and W=w in its consequent is of kind K if and only if, in conjunction with Woodward’s theory of causation, and an appropriate causal model, C implies that V=v is a token cause of W=w.

This fits the cases. Woodward would say that in the absence of arsonists, “The lightning caused the fire” gets to be causal-explanatory by implying “If the lightning had not struck, the field would not have caught on fire.” And the truth of this counterfactual, in the causal model with just variables L and S, is sufficient, in Woodward’s theory, for the lightning to be a cause of the fire. Similarly, Woodward would say that in the presence of the backup arsonist, “The lightning caused the fire” gets to be causal-explanatory by implying

(8) If the lightning had not struck, and the arsonist still not acted, the field would not have caught on fire.

This counterfactual is also of kind K. Let A have value 1 if the arsonist acts, 0 otherwise. Then the causal model of this situation with variable set {L, A, F} has two equations:
A=1 if L=0, 1 otherwise.

F=(L)(1-A).

The directed graph corresponding to this model is as in figure 1. The sequence (L,F) is a path through this graph. Then according to Woodward’s theory of causation, L=1 is a cause of F=1 because, had L=0 and the off-path variable been held at its actual value (so had A still had value 0), F would have had value 0. But this is just the translation of the counterfactual (8) into variable-speak.

Figure 1: directed graph for the causal model of the arsonist scenario

We have seen Woodward explicitly state that (1) is a causal explanation on his theory, and also seen him tell us why. What, then, of my argument that (1) is not a causal explanation on his theory?

5 Against Woodward’s Theory

Suppose Jones knows that the field caught fire, but does not know why. He asks Smith. Smith has learned (1) from a good authority, so knows that the field caught on fire because it was struck by lightning. But he was not himself present at the event, so in particular does not know whether there was an arsonist present. Jones knows this about Smith. (In fact there was an arsonist present.) Smith nevertheless replies to Jones’ question with (1). Since Smith knows that (1) is true, he is in a position to assert it, and since Jones has no reason to doubt Smith’s

---

10 Actually, Woodward’s main focus is on telling us why “The lightning caused the fire” is a causal explanation. This is not the same as (1). But let us assume we are in a context in which (1) entails this causal claim.
testimony, Jones himself comes to know that (1) is true as a result. In summary: upon hearing Smith say (1), Jones came to know why the field caught on fire. He came to know that it caught on fire because it was struck by lightning.

But Smith by his utterance did not exhibit the truth of any counterfactual of kind K concerning the fire. He certainly does not exhibit the truth of the proposition that if the lightning had not struck, the field would not have caught on fire; for this is false. Could he exhibit the truth of the true, and more complicated, counterfactual (8)? It is hard to see how he could. He does not himself know that (8) is true. What’s more, one exhibits the truth of some proposition only if one puts one’s audience in a position to know that it is true. But Smith does not do this. Even if for some strange reason Jones came to believe (8) after hearing Smith utter (1) (and why would he?), this belief would not count as knowledge.

This example shows that Woodward’s theory, on one interpretation, is false. But it also raises an interpretive question. Woodward’s statements (4) and (5) use the word “exhibit.” Here they are again:

(4) [causal] explanation is a matter of exhibiting systematic patterns of counterfactual dependence (191).

(5) the underlying or unifying idea in the notion of causal explanation is the idea that an explanation must answer a what-if-things-had-been-different question, or exhibit information about a pattern of dependency (201).

I have argued that causal explanations do not need to exhibit the truth of certain counterfactuals. But Woodward does not always use the word “exhibit.” In the quotations I discussed in section 4 he uses “imply”; causal explanations must imply the truth of certain counterfactuals. (And I followed this usage in the statement of necessary and sufficient conditions in (7).) Implying and exhibiting are different.

Is the theory better if it requires the implying, rather than the exhibiting, of counterfactuals? No. There are many kinds of implying. Most relevant are two. First is when a speaker
conversationally implies something. Second is when one proposition implies a second, in the sense of its being necessary that if the first is true the second is.

The first sense of “imply” is no help. Smith certainly does not conversationally imply that the counterfactual (8) is true. To do that he would have to himself believe (8), which he does not.¹¹

So what about propositional implication? What if we interpret the requirement as a requirement that causal explanations imply, in this sense, relevant counterfactuals? This case is trickier. (1) does not imply (8). But (1), in conjunction with Woodward’s theory of causation, and the three-variable causal model, does imply (8). Speaking roughly: (1), together with Woodward’s theory, implies that there is some path from L to F relative to which changing L’s value changes F’s; inspection of the model reveals that there is only one such path, and only one alternative to L’s value and F’s value; these are the paths and values that make (8) true. So let us now consider the interpretation of Woodward’s theory got by slightly modifying (7):

(7*) A body of fact constitutes a causal explanation of some event E if and only if that body of fact, in conjunction with Woodward’s theory of causation, and an appropriate causal model, implies the truth of at least one counterfactual of a kind K.

My example is not a counterexample to (7*). But I still maintain that (7*) is false. In fact, my reasons for thinking it false are reasons that Woodward himself should endorse. One complaint he lodges against Hempel’s DN theory of explanation is that it makes the facts in virtue of which a body of fact is explanatory epistemically inaccessible to typical users of that explanation. The DN theory says, roughly, that X explains Y iff (i) X is a conjunction of true propositions, (ii) one conjunct is a law of nature, (iii) X entails Y, and (iv) X wouldn’t entail Y if the law-propositions were removed (Hempel 1965). Example (1) looks like a counterexample to this theory; the explaining-fact, the fact that lightning struck the field, does not contain a law as a conjunct.

¹¹Grice’s original theory of conversational implicature (Grice 1989) entails that X can conversationally imply P only if X believes P.
Hempel replied that (1) still in some sense implicated a law of nature. Woodward objects, in a long discussion in chapter 4 of his book, that this makes the factors in virtue of which (1) counts as an explanation unknown to almost all people who request and offer explanations like (1). Few of them know of any laws that could be added to the proposition that lightning struck to make it entail that the field caught on fire; many of them do not even have the concept of a law of nature.

Woodward’s theory, as stated in (7*), has the same problem. Maybe all causal explanations do imply, in conjunction with Woodward’s theory of causation and some relevant causal model, a counterfactual of kind K. But the counterfactual that is implied is epistemically hidden from at least some of those who request and offer explanations like (1). It is epistemically hidden from Smith and Jones in my example. But they do fine trading in these explanations, even though they do not know which counterfactuals are implied by their explanations.

Woodward does explicitly address the question whether his theory does a better job avoiding the problems he raises for the DN theory. Here is what he says:

singular-causal explanations wear the source of their explanatory efficacy on their face: they explain not because they tacitly invoke a “hidden” law or statement of sufficient (or probabilifying) conditions, but because they identify conditions such that changes in these conditions would make a difference for whether the explanandum phenomenon or some specified alternatives would ensue. The information about such conditions and the counterfactuals associated with them are epistemically accessible and nonhidden. (217)

I have argued that this is wrong. Someone can correctly offer a singular causal explanation like (1) even in a context in which neither she nor her audience has access to the counterfactuals that are associated with (1), the counterfactuals in virtue of which the lightning is a cause of the fire.
Although I think that Woodward’s theory of causal explanation is false, I think its main claims are close to the truth. To get to those truths we need to distinguish between two levels of explanation.

It is easier to start at the higher level and move down. Suppose we have some explanation: Q explains why P.\textsuperscript{12} We can ask of it, why is it true? Why is it that Q explains why P? An answer to this why-question is a higher-level explanation.

A first-level explanation, then, is one where the thing being explained is not itself an explanation. Schematically, “Q explains why P” is a first-level explanation when P does not itself have the form “A explains why B.” Explanations of events, answers to questions of the form “Why did E occur?”, occupy this first level. (The first level does not just contain explanation of events; it also contains, for example, explanations of Galileo’s law of free-fall.)

Woodward’s theory is a theory of causal explanations of events. It therefore concerns explanations at the first level. I have argued that his theory is false. Causal explanations of events need not imply, or exhibit, any relevant counterfactuals.

But there are other explanations that the kinds of counterfactuals Woodward focuses on must appear in. They must appear in certain second-level explanations. If C is a cause of E, then the occurrence of C causally explains why E happened. What explains why \textit{this} is so? Why is it that C causally explains why E happened? It is when answering this question that one must provide a relevant counterfactual. The lightning strike causally explains the fire. Why? If there are no arsonists around, the answer is: because, had the lightning not struck, the field would not have caught on fire. If, on the other hand, there are arsonists around, the answer is different. Then the answer is: because, had the arsonist (still) not acted, and the lightning not struck, the field would not have caught on fire.

According to Woodward’s theory of causation, whenever one event C is a cause of another

\textsuperscript{12}Or: P because Q. I use these interchangeably.
event E, it is because a certain counterfactual (of kind K) relating C and E is true. Woodward holds that the truth of this counterfactual directly (causally) explains why E happened. But that’s not right. instead, what the truth of this counterfactual directly explains is why C is a cause of E, and thereby why C causally explains E. The thing that explains why E happened, that answers the first-level question why E happened, is just the cause C.

I think that Woodward’s mistake is a common one. It appears also, for example, in the DN model. The DN model requires explanations of events to cite laws. But explanations of events do not need to cite laws. Why did Hempel think they did? At one place he answers by sketching a regularity theory of causation: C is a cause of E iff there are other event-types X, Y, and so on, such that it is a law that whenever a C-type event, and X-type event, a Y-type event, and so on, occur, an E-type event occurs (Hempel 1965: 349). But this theory of causation, even if true, does not support the claim that to explain why E happened one must invoke the relevant law. Instead it supports the claim that one must invoke the law to explain why C is a cause of E, or to explain why C explains why E occurred. But natural as it is to think that whenever A explains why B explains C, A also explains C, this thought is false.13

References

Franklin-Hall, Laura (forthcoming). “High-Level Explanation and the Interventionist’s ‘Variables Problem’.” *British Journal for the Philosophy of Science.*


13The false conditional “If A explains why B explains C, then A explains C” should not be confused with the conditional “If A explains B and B explains C, then A explains C.” This second conditional is the claim that explanation is transitive. This is not implausible, but is not the I am rejecting. (Not everyone accepts that explanation is transitive; those who deny that causation is transitive will not accept it.)


