Would-Cause Semantics
Author(s): Phil Dowe
Reviewed work(s):
Source: *Philosophy of Science*, Vol. 76, No. 5 (December 2009), pp. 701-711
Published by: The University of Chicago Press on behalf of the Philosophy of Science Association
Stable URL: http://www.jstor.org/stable/10.1086/605796
Accessed: 12/01/2012 01:18

Would-Cause Semantics

Phil Dowe[†‡]

This article raises two difficulties that certain approaches to causation have with would-cause counterfactuals. First, there is a problem with David Lewis's semantics of counterfactuals when we 'suppose in' some positive event of a certain kind. And, second, there is a problem with embedded counterfactuals. I show that causal-modeling approaches do not have these problems.

1. Introduction. Possible causation is an important but overlooked topic in the study of causation. Of particular significance are 'would-cause counterfactuals' such as

1. 'Had an event *c* of kind *C* occurred, *c* would have caused an event *e* of kind *E*', where actually no events of kinds *C* or *E* occur.

Would-cause counterfactuals are important in their own right for both common sense and science. 'If I had called Mum it would have caused her happiness' is an instance of 1, and such homely would-cause claims can be generated indefinitely (see Dowe 2009, Section 2). Scientific examples are also ubiquitous: 'had we applied radiation in time that would have killed the tumor'. In this article, I raise two difficulties that David Lewis's semantics for counterfactuals have with would-cause counterfactuals. Before turning to these difficulties, I survey some theoretical uses of would-cause counterfactuals.

2. Theoretical Uses of Would-Cause Counterfactuals. The first example of the use of would-cause counterfactuals is Lewis's theory of contrastive causal explanation. According to Lewis (1986), we explain why *e* rather than *e** by supplying details of the actual causal history of *e* that differentiate it from the counterfactual causal history of *e**:

[†] To contact the author, please write to: Philosophy, University of Queensland, Brisbane, Queensland 4072, Australia; e-mail: p.dowe@uq.edu.au.

[‡] I would like to thank Stephen Barker and Jonathan Schaffer for comments on some of this material.

Philosophy of Science, 76 (December 2009) pp. 701–711. 0031-8248/2009/7605-0009$10.00

> Why did I visit Melbourne in 1979, rather than Oxford or Uppsala or Wellington? Because Monash invited me. That is part of the causal history of my visiting Melbourne; and if I had gone to one of the other places instead, presumably that would not have been part of the causal history of my going there. It would have been wrong to answer: Because I like going to places with good friends, good philosophy, cool weather, and plenty of trains. That liking is also part of the causal history of my visiting Melbourne, but it would equally have been part of the causal history of my visiting any of the other places, had I done so. (Lewis 1986, 229–230)

Call the details of the actual causal history of e c_1, c_2, c_3, ... Call the details of the counterfactual causal history of e^* c_1^*, c_2^*, c_3^* . . . . Then Lewis's account of causal explanation appeals to claims like

> c_i caused e, and had c_i^* occurred, it would have caused e^*.

The latter is a would-cause counterfactual of the kind we are interested in. According to Lewis, you can plug any theory of causation into his theory of causal explanation, but of course we would be interested in Lewis's theory of causation. Actually, Lewis has several theories of causation; let us take his original theory. According to Lewis (1973), *c* causes *e* if *c* and *e* are distinct events that occur and 'had *c* not occurred, *e* would not have occurred' is true. This latter is true if and only if (iff) *e* does not occur in any of the closest worlds in which *c* does not occur and false if it does. This evaluation of the counterfactual is to 'suppose away' a positive event, let us say. But to evaluate the would-cause counterfactual, we need, first, to, let us say, 'suppose in' some positive event of a certain kind. And, second, we need embedded counterfactuals since we need to analyze the cause in the would-cause counterfactual via the counterfactual theory of causation. It is a reasonable demand, then, that Lewis's semantics can handle both these things.

The second example is my own account of 'causation' by absences (Dowe 2000, 132–139), according to which absences cannot be causes or effects, and 'causation' involving negatives is not genuine causation but a stand-in that I call quasi causation, which essentially involves counterfactuals about genuine causation. Quasi causation supervenes on patterns of actual and possible genuine causation (see also Armstrong 2004, 66–67). I will just summarize three well-known kinds (for more, see Dowe 2000, 132–139), expressed in terms of events. For a discussion of the quasi-causal relata, see Dowe 2009.

a) Prevention by omission: not-*a* quasi caused not-*b* if neither *a* nor *b* occurred, and
   1. if *a* had occurred, *a* would have caused *b*.

b) Prevention: *a* prevented *b* if *a* occurred and *b* did not, and there occurred an *x* such that
   (P1) there is a causal interaction between *a* and the process due to *x*, and
   (P2) if *a* had not occurred, *x* would have caused *b*.
c) Omission: not-*a* quasi caused *b* if *b* occurred and *a* did not, and there occurred an *x* such that
   (O1) *x* caused *b*, and
   (O2) if *a* had occurred then *a* would have prevented *b* by interacting with *x* where prevention is analyzed as above.

Example a1 is pretty much the would-cause counterfactual 1. Examples b and c involve more complex kinds of possible causation. First, note that a and c involve counterfactuals that suppose in a positive. And second, to analyze 'prevents' in c, one appeals to b, which itself involves a counterfactual clause P2. Thus, this account requires embedded counterfactuals. Examples a and b do contain 'cause' within the counterfactual, but this in itself does not require embedded counterfactuals unless one wants to use a counterfactual theory of causation. My own account of genuine causation (Dowe 2000) does not. In discussing this account of absence causation (2001, 221), I said "it's B.Y.O. semantics," meaning that any semantics of counterfactuals should work. That was a mistake. In the light of Section 3 (below), I should have said "it is D.I.Y. semantics."

The third example is Lewis's account of absence causation as a *variety* of causation (2004b, 284–285). Lewis gives the name 'biff' to the intrinsic relation between distinct events that is typically associated with chance raising (cf. Menzies' 1996 account of causation). Then

i) Event *c* directly causes event *e* iff *c* stands to *e* in the relation that occupies the biff role, or, for short, iff *c* biffs *e*.
ii) The absence of any event of kind *C* directly causes event *e* iff, had there been an event *c* of kind *C*, *c* would or might have biffed some event *d* incompatible with event *e*.
iii) Event *c* directly causes the absence of any event of kind *E* iff *c* biffs some event *d* incompatible with any event of kind *E*.
iv) The absence of any event of kind *C* directly causes the absence of any event of kind *E* iff, had there been an event *c* of kind *C*, *c* would or might have biffed some event *e* of kind *E*. (Lewis 2004b, 284–285)

This defines direct causation, and indirect causation is the ancestor of direct causation.

Since biff is a variety of causation, iv is pretty much my would-cause counterfactual 1. First, note that ii and iv involve counterfactuals that suppose in a positive. And second, ii–iv all involve counterfactuals about biff. The 'biff role' is a chance-raising role. Chance raising involves counterfactual chances: *c* biffs *e* only if the actual chance of *e* is greater than it would be had *c* not occurred. Thus, to evaluate counterfactuals about biff, we need embedded counterfactuals. Again, it is reasonable to demand that Lewis's semantics for counterfactuals can handle both these things.

Finally, there are other theories that do not employ would-cause counterfactuals but that do involve counterfactuals that suppose in positives. These will be open to the first but not the second of the two difficulties discussed in the next section. For example, Lewis (2004a) apparently gives a different account of absence causation to the one discussed above. On this account, 'not-*a* caused not-*b*' obtains on account of the true proposition 'had *a* occurred then *b* would have occurred', although this does not indicate a literal causal relation. This supposes in a positive. We would have hoped that Lewis's semantics for counterfactuals can handle this.

According to the contrastivist theory of causation (Maslen 2004; Schaffer 2005; Northcott 2008), *c* rather than $\boldsymbol{C}^*$ causes *e* rather than $\boldsymbol{E}^*$ iff (i) *c* and *e* are actual distinct events, (ii) $\boldsymbol{C}^*$ is a set of possible events (situations) alternative to *c* and $\boldsymbol{E}^*$ is a set of possible events (situations) alternative to *e*, and (iii) for every event c_i^* in $\boldsymbol{C}^*$ there is an event e_i^* in $\boldsymbol{E}^*$ such that if c_i^* had happened then e_i^* would have happened. This does not require embedded counterfactuals. But even for core examples of causation, counterfactuals that suppose in positives must be evaluated. Maslen, Shaffer, and Northcott each appeal to Lewis's semantics for counterfactuals, without strongly endorsing the approach. It seems necessary for this theory, then, that Lewis's semantics work for such counterfactuals.

3. Two Problems with Lewis's Semantics. Although Lewis's semantics (1979) remain by far the most used and discussed semantics in the context of counterfactual theories of causation, many difficulties with that account have been identified (e.g., Barker 1999; Elga 2000). The current concern is with difficulties that apply not to the evaluation of 'had *c* not occurred, *e* would not have occurred' but with the extension of the approach to 1. I raise two problems. The first concerns the general problem of supposing in positives, and the second concerns embedded counterfactuals.

3.1. Supposing in Positives. When *c* does not actually occur, what is (are) the closest *c* world(s)? (More precisely, when no event of kind *C* occurs, what is [are] the closest world[s] in which some event *c* of kind *C* occurs?) To adapt an example due to Hall (2002), suppose Billy shoots down Enemy's plane before Enemy can shoot down Suzy. Suzy goes on to successfully bomb the target. Suppose also that Enemy's being shot down causes a grass fire. Grant the counterfactuals 'had Billy not shot down Enemy, Enemy would have shot down Suzy' and 'had Enemy shot down Suzy he would have prevented the bombing'. To evaluate 'had Enemy shot down Suzy . . . ', we want the closest world in which Enemy shot down Suzy. Lewis's semantics say that, first, we must avoid large miracles (big, widespread, and diverse miracles); second, we maximize regions of perfect match with actuality; and, third, we minimize small, local, simple miracles (1979). The idea is that the closest Enemy-shoots-Suzy world is a world (or worlds) with perfect match until a time just before a small transition miracle brings about Enemy's shooting down of Suzy, and then no further miracles occur.

First, there are worries about backtracking. When we make the claim 'had Enemy shot down Suzy he would have prevented the bombing', plausibly we are supposing Enemy never was shot down by Billy. Canonical backtracking counterfactual reasoning is where we suppose away an effect and make inferences about the absence of its (usually earlier) cause (Lewis 1979, 33); Lewis's semantics are not intended for backtrackers. In our case we do not make inferences about what would have been the cause of Enemy's shooting down Suzy, but we do seem to incorporate assumptions about the causes into the antecedent: 'had Enemy shot down Suzy (having not been shot down by Billy) he would have prevented the bombing'. But even if this does not count as backtracking, it still completely undermines the point of Lewis's semantics. Including this much of the causal past in the antecedent will make false counterfactuals true, such as 'had Enemy shot down Suzy (having not been shot down by Billy) he would have prevented the grass fire'. One point of maximizing regions of perfect match is to rule out such spurious claims.

Second, supposing that we can avoid assuming this background explicitly in the antecedent, there are worries about whether the miracle it would take to have Enemy shoot down Suzy could count as a small miracle. The miracle cannot be one such that Enemy and his plane—or an Enemy-like person flying a replica of Enemy's plane—appear ex nihilo *while at the same time* Enemy and his plane lie scattered across the countryside. This would violate all plausible theories of transworld and transtemporal identity. No, to get Enemy into place to shoot down Suzy, we need (in a short transition time) to reassemble the scattered parts of Enemy and his plane, which would involve numerous independent miracles. According to Lewis, a big miracle is a "big widespread diverse violation of law," which is composed of small miracles, but these are (1) many, (2) not jointly localized, and (3) diverse, that is, involving violations of different sorts of laws (Lewis 1979, 47). All three features are true of our divergent miracle, hence it is a big miracle. Therefore, a closer world is one that sacrifices some perfect match to avoid the large miracle: Billy's gun jams by a single small miracle, and Enemy is not shot down but instead goes on to shoot down Suzy. But again this opens the door to spurious claims. If the closest world in which Enemy shoots down Suzy is a world in which the only miracle is that Billy's gun jams, then 'had Enemy shot down Suzy he would have prevented the grass fire' comes out as true. So on the counterfactual theory of causation, Enemy's not shooting down Suzy causes the grass fire; on the quasi-causation theory, it is a quasi cause.

Supposing in a positive event does not always involve big miracles. Former prime minister Howard's failure to say 'sorry' for the generation of aboriginal children stolen from their parents could be supposed away by inserting a mere pang of conscience. But many other cases look much more like big miracles: in evaluating 'had the doctor operated, the patient would have been saved' in supposing away the doctor's failure to operate, we need to get the doctor from his golfing holiday to the operating theater and the patient from her shopping holiday in Paris and into the operating theater. This requires distant diverse miracles on the golf course, in Paris, and in the hospital. Again, avoiding big miracles is more important than maximizing regions of perfect match, so the closest world in which the doctor operates might be one in which he was never invited on the golfing holiday in the first place. Again, this opens the door to spurious causal claims.

Finally, another difference between supposing away and supposing in positive events is that in the former case the spatiotemporal location of the desired small miracle is pretty much fixed: a small time before the occurrence of the positive. Not so with absences—there is much leeway for where and when the positive occurs. In fact, given the requirement to maximize regions of perfect match, the closest world in which the doctor does operate is one in which he operates too late to save the patient—thus we need to include a rider such that 'had the doctor operated' is read as 'had the doctor operated in due time'. But still, even granting this (as perhaps we should), the closest worlds when supposing away an absence are last-minute worlds—for example, the doctor performs the operation later rather than earlier. Perhaps that is just an unintended curiosity.

It is not difficult to find scientific examples of this problem. 'Had Mars collided with Earth at the same time as comet *x* collided with Earth then . . . .' It would take a large miracle to transition from the actual course of the solar system to the supposed collision with Mars, if this is to occur in a relatively small transition time just before the actual collision with comet *x*. Since avoiding big miracles is more important than maximizing regions of perfect match, the closest world in which the collision with Mars occurs is one with a small miracle that brings about a much earlier variation to the orbit of Mars. But then other effects of the earlier varied trajectory of Mars will count as effects of the collision between Earth and Mars.

3.2. Embedded Counterfactuals. Consider the analysis of cause in 1. On the counterfactual theory of causation 'had *c* occurred, *c* would have caused *e*' comes out as 'had *c* occurred, then had *c* not occurred then *b* would not have occurred'. We assume, actually, not-*c*, so from the closest worlds in which *c* occurs, all the closest worlds in which ~*c* holds need to be ~*e* worlds. There is no guarantee that this will covary with our intuitions (see Barker 2009). The reason is that on Lewis's semantics, to suppose *c* requires a miracle. This is a violation of the laws of nature at the actual world but not at those *c* worlds. The laws of nature at a world for Lewis are the best system analysis of the distribution of particulars at that world. The laws at the closest *c* worlds may then be different from those of the actual world. To suppose, from any of those worlds that had ~*c*, requires analysis in terms of the laws at the *c* world. Thus, whether ~*e* obtains depends in general on the different laws and will in general give different answers. Thus, Lewis's semantics cannot handle embedded counterfactuals. So the counterfactual theory of causation with Lewis's semantics cannot handle would-cause counterfactuals.

For example, suppose I do not actually drink arsenic, but I do drink water, and we are interested in the counterfactual 'had I drunk arsenic it would have caused my death', which we think is true. Suppose by Lewis's semantics that the closest arsenic-drinking world W_1 is one in which a small miracle turns the water I am about to drink into arsenic. Then at that world, the laws allow for water to turn into arsenic. So, suppose further that, in the actual world W_a, just before my drinking, the water is in a state S (involving a very particular combination and concentration of various minerals, say) that as it happens has never previously existed in W_a. At W_1, the small miracle is S to arsenic. Suppose that the laws of W_1 mandate the transition S to arsenic. From W_1, suppose the closest not-drinking-arsenic worlds are those with a perfect match with W_a and W_1 up to the time of the S-to-arsenic transition and in which by a small miracle (according to the W_1 laws) the S-to-arsenic transition fails to occur. At one of those worlds it happens that the water in my stomach is in state S and turns to arsenic and I die anyway. Closest-world analysis requires that I do not die in any of the closest worlds, so it is not true that 'had I drunk arsenic it would have caused my death'.

The desiderata for a semantics capable of handling 1 are clear enough. For a minimal correction to Lewis's semantics, we need (*a*) to distinguish in a noncircular way those big miracles that produce multiple partial causes of an event we are supposing in from those big miracles that wipe out the multiple traces of an event we are supposing away and (*b*) to appeal to the actual laws in evaluating embedded counterfactuals. The latter approach has its advocates (e.g., Barker 1999). I know of no account that meets the former desideratum.

4. Alternative Approaches. There are alternative counterfactual approaches to causation that explicitly eschew Lewis's semantics and, in particular, Lewis's appeal to miracles. Most prominent is the causal-modeling approach typified by Woodward (e.g., 2003). I will follow Craver's (2007) accessible account, which draws on Woodward (2003) and Woodward and Hitchcock (2003), who in turn draw on the causal-modeling tradition of Pearl (2000) and Spirtes, Glymour, and Scheines (2000). According to Craver, "Variable X is a cause of variable Y in conditions W, if and only if it is possible in conditions W to change the value of Y with an ideal intervention that changes the value of X" (2007, 94), where

> An *ideal* intervention I on X with respect to Y is a change in the value of X that changes Y, if at all, *only via* the change in X. More specifically, this requirement implies that:
> (I1) I does not change Y directly;
> (I2) I does not change the value of some causal intermediate S between X and Y except by changing the value of X;
> (I3) I is not correlated with some other variable M that is a cause of Y; and
> (I4) I acts as a "switch" that controls the value of X irrespective of X's other causes, U. (Craver 2007, 96)

This account is not concerned at all with ranking worlds according to miracles and perfect match. Miracle semantics are replaced by the notion of an ideal intervention. I do have worries about this account, but I do not want to canvass those here. Rather, I want to show that the account does avoid the two problems I have raised for Lewis's semantics.

4.1. Supposing in Positives. The first problem concerns the closest world when we suppose in a positive. The proponents of causal-modeling claim that they can smoothly handle absence causation (e.g., Woodward 2003; Craver 2007, 104). I will show that they do avoid the problems with supposing in positives that I have claimed besets Lewis's semantics, precisely because they do not appeal to such a similarity measure on worlds. Take the claim 'Enemy's not shooting down Suzy caused the bombing' (causation by omission). Let

$X = 1$ if Enemy shoots down Suzy, 0 if he does not;

$Y = 1$ if Suzy bombs target, 0 if she does not; and

$U = 1$ if Billy shoots Enemy, 0 if he does not.

Given that Billy actually does shoot down Enemy, the relevant laws (with an assumed direction) can be encoded as equations:

$$U = 1,$$
$$X \simeq U,$$
$$Y \simeq X.$$

We need an intervention I to bring about X = 1. This then replaces the second equation (see, e.g., Hitchcock 2007), while holding fixed the actual value of U. The equations become

$$U = 1,$$
$$X = 1,$$
$$Y \simeq X.$$

Thus Y = 0, and we do indeed have a case of causation (by omission). The important point, though, is that we allow I if it fits I1–I4 above, which requires holding fixed U. For causal modeling it does not matter that intervention I might count as a large miracle on Lewis's story. And there is no ranking of large miracles versus perfect match that would entail that we should, for example, vary U to vary X. Thus, causal modeling does not lead to spurious causal claims in the way Lewis's semantics do.

4.2. Embedded Counterfactuals. The second problem concerns the embedded counterfactuals we get when we analyze cause in 1, on a counterfactual theory of causation. The version of causal modeling under consideration is a counterfactual theory, but it does not run into the same embedding problem discussed above. Take the would-cause counterfactual 'had Enemy shot down Suzy he would have prevented the bombing'. Assume for the sake of the argument that prevention is causation. Causal modeling is not a method for dealing with counterfactuals in general, and to my knowledge its proponents have not turned their attention to would-cause counterfactuals. But it is not difficult to envisage how that might go.

Recipe: Vary the independent variables from actuality to get the cases in which the antecedent is true, in our case, where X = 1. Then ask, does X = 1 cause Y = 0? If it does, in all cases in which X = 1, the would-cause counterfactual is true. We get

$$U = 0,$$
$$X \simeq U,$$
$$Y \simeq X.$$

Then test by an ideal intervention on X, for example, Enemy's gun jams, so replace the second equation with X = 0:

$$U = 0,$$
$$X = 0,$$
$$Y \simeq X.$$

Then Y = 1, so X = 1 does cause Y = 0. Thus, we cannot end up with different laws for evaluating the cause in the counterfactual scenario. Unlike Lewis's account, causal modeling, on its own terms, has no problem with embedded counterfactuals when evaluating would-cause counterfactuals.

REFERENCES

Armstrong, David (2004), *Truth and Truthmakers*. Cambridge: Cambridge University Press.

Barker, Stephen (1999), "Counterfactuals, Probabilistic Counterfactuals, and Causation", *Mind* 108: 427–469.

——— (2009), "Contrary to Fact", unpublished manuscript. Nottingham: University of Nottingham.

Craver, Carl (2007), *Explaining the Brain*. Oxford: Oxford University Press.

Dowe, Phil (2000), *Physical Causation*. New York: Cambridge University Press.

——— (2001), "A Counterfactual Theory of Prevention and 'Causation' by Omission", *Australasian Journal of Philosophy* 79: 216–226.

——— (2009), "Absences, Possible Causation, and the Problem of Non-locality", *Monist* 92 (January): 23–40.

Elga, Adam (2000), "Statistical Mechanics and the Asymmetry of Counterfactual Dependence", *Philosophy of Science* 68: S313–S324.

Hall, Ned (2002), "Non-locality on the Cheap?", *Noûs* 36: 276–294.

Hitchcock, Christopher (2007), "Prevention, Preemption, and the Principle of Sufficient Reason", *Philosophical Review* 116: 495–532.

Lewis, David (1973), "Causation", *Journal of Philosophy* 70: 556–567.

——— (1979), "Counterfactual Dependence and Time's Arrow", *Noûs* 13: 455–476.

——— (1986), "Causal Explanation", in his *Philosophical Papers*. Vol. 2. Oxford: Oxford University Press, 214–240.

——— (2004a), "Causation as Influence", in John Collins, Ned Hall, and Laurie Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 75–106.

——— (2004b), "Void and Object", in John Collins, Ned Hall, and Laurie Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 277–290.

Maslen, Cei (2004), "Causes, Contrasts, and the Nontransitivity of Causation", in John Collins, Ned Hall, and Laurie Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 341–357.

Menzies, Peter (1996), "Probabilistic Causation and the Preemption Problem", *Mind* 105: 85–117.

Northcott, Robert (2008), “Causation and Contrast Classes”, *Philosophical Studies* 139: 111–123.

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

Schaffer, Jonathan (2005), “Contrastive Causation”, *Philosophical Review* 114: 297–328.

Spirtes, Peter, Clarke Glymour, and Richard Scheines (2000), *Causation, Prediction, and Search*. Springer Lecture Notes in Statistics, 2nd rev. ed. Cambridge, MA: MIT Press.

Woodward, James (2003), *Making Things Happen*. New York: Oxford University Press.

Woodward, James, and Christopher Hitchcock (2003), “Explanatory Generalizations”, pt. 1, “A Counterfactual Account”, *Noûs* 37: 1–24.