6

The Counterfactual Theory

Chapter 3 explored the possibilities of explaining the asymmetry of causation in terms of the relations between causation and time. Chapter 5 explored the possibilities of generating a theory of causal asymmetry from the connections between causation and agency. This chapter explores another attractive set of connections – those between causation and counterfactuals. When one asserts that one event *a* causes another distinct event *b*, then it seems that one is committed to the counterfactual: "If *a* had not occurred, then *b* would not have occurred either." Hume himself wrote ". . . we may define a cause to be *an object, followed by another, and where all the objects similar to the first, are followed by objects similar to the second*. Or in other words, *where if the first object had not been, the second had never existed*" (*Inquiry* p. 51). The "other words" here are indeed "other words." Hume seems mid-definition to leap from a regularity to a counterfactual theory of causation. This chapter explores the possibility of constructing a counterfactual theory of causation.

6.1 Lewis's Theory

According to David Lewis, if *a* and *b* are distinct events that actually occur, then *b* causally depends on *a* if and only if, if *a* were not to occur, then *b* would not occur either (1973a). If the match had not been struck, then it would not have lit. In Lewis's usage (in contrast to mine), "causal dependence" is a relation among token events, which is sufficient for causation. It is not the same thing as causation, because in cases of preemption it is not transitive. Lewis takes causation to be the ancestral of causal dependence: *a* causes *b* if and only if *b* causally depends on *a* or there is a chain of causal dependence linking *a* and *b*. One should not take causal dependence to be "direct causation." The collapse of the Soviet Union may have causally depended on Lenin's death seventy years earlier.

Since I am postponing considering problems concerning preemption until chapter 13, I shall simplify Lewis's theory and take his account of causal dependence to be an account of causation. This begs no questions and eases the comparisons between his theory and the accounts discussed in previous chapters.[1] I shall formulate this simplification of Lewis's theory as:

> **L** (*Lewis's theory*) *a* causes *b* if and only if *a* and *b* are distinct events and if *a* were not to occur, then *b* would not occur either.

I am taking the occurrence of *a* and *b* as implicit in the claim that they are distinct events. Let us say that *b* is *counterfactually dependent* on *a* if and only if both the following counterfactuals are true: "If *a* were to occur, then *b* would occur" and "If *a* were not to occur, then *b* would not occur" (Lewis 1973a, p. 166). If *a* and *b* both occur, then the first counterfactual is automatically true, since the closest possible world in which *a* occurs is the actual world and in that world *b* occurs too. **L** can be restated as the claim that counterfactual dependence among distinct events is necessary and sufficient for causation. Like Lewis I shall take counterfactuals to be statements that may be true or false.

To evaluate the counterfactual, "if *a* were not to occur, then *b* would not occur," one considers "possible worlds" in which *a* does not occur. Some of these possible worlds will be more similar to the actual world than are others. The counterfactual is true if some possible world without *a* (some "non-*a* possible world") in which *b* does not occur is more similar to the real world than any non-*a* possible world in which *b* occurs. In Lewis's theory, understanding of counterfactuals in this way grounds understanding of causality.

The "comparative overall similarity" of possible worlds is a vague relation, and Lewis argues that we should fill it in so as to fit our intuitions concerning which counterfactuals are true. When Lewis does this, he finds that the comparative overall similarity of possible worlds depends on several factors (Lewis 1979, 1986c). Widespread and diverse differences between the laws of nature count against similarity, while "perfect match" over a time interval counts heavily for similarity. Small differences between laws of nature are relatively unimportant. These appear from the perspective of the actual world as localized and minor miracles. Possible worlds that differ from the actual world only with respect to small numbers of such miracles are very similar to the actual world. In Lewis's view, the closest non-*a* possible world should have exactly the same history as the actual world until shortly before the time at which *a* occurs in the actual world. If *a* is causally determined, there will then be some relatively isolated violation of the laws of the actual world – so that *a* fails to occur. If the most similar possible worlds had the same laws of nature and determinism were true, the most similar possible worlds would have to have completely different histories. Permitting "miracles" avoids this absurdity. When one considers what would be true if *b* had not occurred, one holds fixed the past

[1] Taking causation to be the ancestral of causal dependence also allows *c* to be a cause of *e* when the influence of *c* on *e* along one path cancels out its influence along another path. See §12.2.

(including the occurrence of its cause, *a*) and “gets rid” of *b* by means of a “small miracle” just before *b* occurs.

Laws of nature are not sacrosanct in evaluating counterfactuals. Since perfect match counts so heavily in determining overall similarity among possible worlds, the closest possible world without the effect, *b*, will not diverge from the actual world until after the cause, *a*, occurs. So causes are not counterfactually dependent on their effects. Similarly, effects *e* and *f* of a common cause *c* are not counterfactually dependent on one another, because possible worlds without *e* that diverge from the actual world only after *c* has occurred will be closer to the actual world than are possible worlds without *c*. Thus Lewis purports to solve what he calls “the problem of effects” and “the problem of epiphenomena.” Lewis argues that one must resist the temptation to say that causes are sometimes counterfactually dependent on their effects and that effects of a common cause are sometimes counterfactually dependent on one another.

This account of the asymmetry of causation might appear arbitrary. If perfect match between possible worlds is so important in considering their similarity, why isn’t the most similar world without *a* one in which *a* fails to happen owing to one miracle and in which the future then goes on just the same as in the actual world (i.e., includes *b*), owing to another miracle? One might also ask why a possible world with just the same past and then, owing to a miracle, a different future is more similar to the actual world than a possible world with a different past and then, owing to a miracle, the same future.

Lewis’s answer is an empirical one: A single small miracle will not erase the consequences of *a*’s failure to occur, nor will a single miracle lead a possible world with a different past to have just the same future. There is an asymmetry of miracles, which results from an asymmetry of overdetermination (1979, pp. 48–51). Causes leave many traces, which require many miracles to erase. The fact that the most similar possible worlds are just like ours until just before an event fails to occur is not an arbitrary stipulation, and it does not presuppose that causes must precede their effects. On the contrary, facts about overdetermination coupled with a correct understanding of the semantics for counterfactuals imply that the most similar possible worlds involve miracles just before the supposed occurrence or nonoccurrence. §6.5* shows that the asymmetry of overdetermination is a consequence of the independence condition (**I**) when causation is deterministic.

It is undeniable that people sometimes say that if an event had failed to occur, then one or another of its causes must have failed to occur. Lewis responds that these counterfactuals involve a nonstandard “backtracking” interpretation. Given the standard understanding of counterfactuals, such claims are, Lewis maintains, false. This response still seems ad hoc, and I shall argue that it is unnecessary.

6.2 Asymmetry Without Miracles

There is another way in which a defender of **L** might respond to the problem of effects and to the problem of epiphenomena. Consider exactly what Lewis writes:

> [1] The proper solution to both problems [of effects and of epiphenomena], I think, is flatly to deny the counterfactuals that cause the trouble. [2] If *e* had been absent, it is not that *c* would have been absent. . . . [3] Rather, *c* would have occurred just as it did but would have failed to cause *e*. [4] It is less of a departure from actuality to get rid of *e* by holding *c* fixed and giving up some or other of the laws and circumstances in virtue of which *c* could not have failed to cause *e*, rather than to hold those laws and circumstances fixed and get rid of *e* by going back and abolishing its cause *c*. (1973a, p. 170)

Focus particularly on the second and third sentences in this quotation. The second denies the counterfactual, $\sim O(e) > \sim O(c)$, while the third asserts the counterfactual, $\sim O(e) > O(c)$ (where "$O(e)$" is the proposition that *e* occurs, and ">" represents the counterfactual conditional). If the law of the conditional excluded middle holds, then the second sentence entails the third. (The law of the conditional excluded middle states: $\mathbf{P} > \mathbf{Q}$ or $\mathbf{P} > \sim\mathbf{Q}$. See Lewis (1973b), pp. 79–82.) But Lewis rightly denies the law of the conditional excluded middle and may deny $\sim O(e) > \sim O(c)$ without asserting, "If *e* had been absent, . . . *c* would have occurred just as it did." In order to account for the asymmetry of causation, Lewis need only deny the counterfactual dependence of specific causes on their effects. He does not need to assert that specific causes would still occur if their effects failed to occur.[2]

This suggests a way to defend **L** without invoking last-minute miracles. Indeed Lewis's wording in the last sentence in the quotation suggests that miracles between the cause and effect one is concerned with may not be necessary. He writes, "It is less of a departure from actuality to get rid of *e* by holding *c* fixed and giving up some or other of the laws *and circumstances* in virtue of which *c* could not have failed to cause *e* . . . [my italics]." If *e* had not occurred, then possible worlds without *c* are no more (or less) similar to the actual world than are possible worlds without some other cause or causal condition of *e*.

Suppose that *e* has two causes, *a* and *c*, and consider the three counterfactuals:

1. If *e* had not occurred, either *a* or *c* would not have occurred.
2. If *e* had not occurred *a* would not have occurred.

[2] One must thus reject Michael McDermott's revision of Lewis's theory (1995, p. 137), which says that if *c* and *e* occur, then *e* causally depends on *c* if and only if if *c* had not occurred, then *e* *might* not have. For the negation of such a might counterfactual (with the nonoccurrence of an effect as its antecedent) is identical to the assertion that if the effect had not occurred, the cause would have occurred just the same.

3. If *e* had not occurred *c* would not have occurred.

(1) does not imply (2) or (3). If non-*e* possible worlds without *a* are no more or less similar to the actual world than are non-*e* possible worlds without *c*, then both (2) and (3) are false. There is no need to deny (1). These claims are not epistemic. The problem is not that we are unable to *find out* whether (2) or (3) is true: If non-*e* possible worlds without *a* are no more or less similar to the actual world than are non-*e* possible worlds without *c*, then both (2) and (3) are false.

If there are such ties in similarity, Lewis could agree that if an effect determined by its causes were not to occur, then some of its causes wouldn't have occurred without having to assert that any one of its causes failed to occur.[3] He could still deny claims such as "if *e* had not occurred, then *a* would not have occurred." And this denial is all that a counterfactual theory needs in order to account for the asymmetry of causation. The solution to the problem of epiphenomena is similar. Lewis need not maintain that if one effect of a common cause failed to occur, then the others would still occur. He need only deny all specific counterfactual dependence of one effect of a common cause on another.

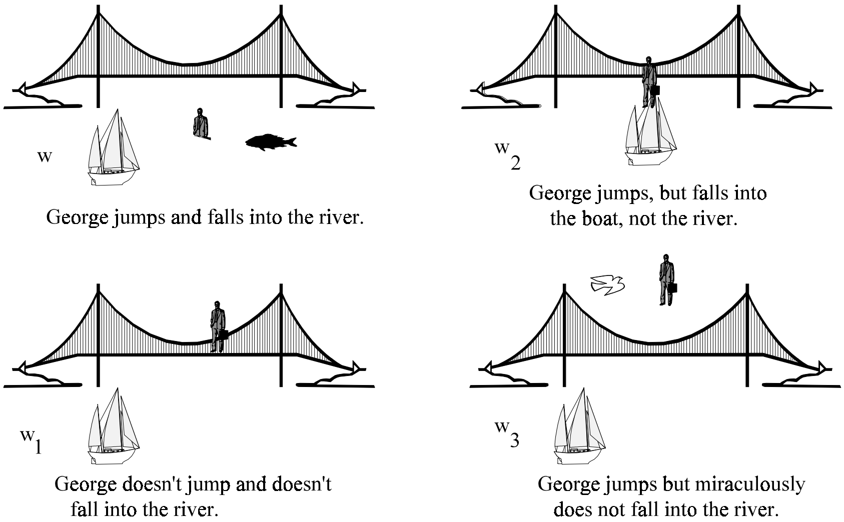
6.3 Why Causes Are Not Counterfactually Dependent on Their Effects

A specific cause *a* of an event *b* will not be counterfactually dependent on *b*, if *b* has other independent causes, and there are non-*b* possible worlds without one of *b*'s other causes that are at least as similar to the actual world as are non-*b* possible worlds without *a*.[4] If the independence condition of chapter 4 holds, then there will be a multiplicity of independent causes. It is not particularly plausible to maintain that there will always be ties among the closest possible worlds without one or another of these causes, but the implausibility of this claim is no worse than the implausibility of Lewis's views on similarity among possible worlds.

For example, suppose George jumps off the Brooklyn Bridge and plunges into the East River. It is plausible to claim that if George had not plunged into the river, then he would not have jumped. People are inclined to find not jumping overwhelmingly the *most likely explanation* for George's

[3] As Christopher Hitchcock pointed out to me, to accept the revised version of his theory, Lewis would have to give up his account of causal preemption. Since that account fails in any event in cases of so-called late preemption (1986d, pp. 193–212) and since a simpler and more powerful solution to the problem of preemption is available (see §13.2), this seems no great loss.

[4] Artificial compound events create problems here, which are discussed on pp. 135–6.



George jumps and falls into the river.

George jumps, but falls into the boat, not the river.

George doesn't jump and doesn't fall into the river.

George jumps but miraculously does not fall into the river.

Figure 6.1: George jumps off the Brooklyn Bridge

hypothetical failure to plunge into the East River.[5] Lewis insists that counterfactual theorists need not be governed by such intuitions. Instead one needs to find an interpretation of similarity that will not imply that George's jumping is counterfactually dependent on his falling in the river, since his jumping is obviously not causally dependent on his falling into the river.

Let *w* in figure 6.1 be the actual world, in which George jumps from the Brooklyn Bridge and plunges into the East River. Among the causes of his plunging into the East River are his jumping and, let us suppose, a boat being downstream of the bridge rather than beneath him. Consider then the following three possible worlds (shown in figure 6.1) in which George does not plunge into the East River. In w_1 he does not jump. In w_2 he jumps, and the boat is beneath the bridge, so that he falls onto it. In w_3, he jumps, the boat stays put, but by some miracle he does not plunge into the river. Perhaps he is blown to shore. Most people would judge worlds like w_1 to be more similar to the actual world than are worlds like w_2 or w_3. If causal dependence is counterfactual dependence, Lewis had better not agree. So it had better not be the case that the miracle that moves the boat is, like the

[5] In an incautious moment, Lewis himself writes that we should say that if a barometer had read higher, it would have been malfunctioning rather than that the pressure would have been higher. His reason is that "The barometer, being more localized and more delicate than the weather, is more vulnerable to slight departures from actuality" (1973a, p. 169). If one takes this literally, then Lewis is saying that the functioning of the barometer is counterfactually dependent on the reading, and one has a surprising example of backwards causation. Such remarks do not reappear in Lewis's later discussion in "Counterfactual Dependence and Time's Arrow," (1979) and I shall assume that they are a mistake.

miracle that somehow keeps George out of the river in w_3, a "big" one, or that it has to begin earlier than the miracle that prevents the jumping. For then Lewis would have to agree that w_1 is more similar to w than are w_2 or w_3, and his account would falsely imply that George's plunging into the river causes his jumping.

There are two ways to avoid this result. One is to hold that w_3 is more similar to w than are w_1 or w_2. I would not do so, because I deny that worlds with miracles between cause and effect are more similar to the actual world than are worlds without such miracles. Lewis permits miracles between cause and effect, but I doubt that he would want to rely on one here, because the miracle that keeps George out of the water in w_3 does not seem to be small and isolated. The only other way to avoid the counterfactual dependence of the jumping on the falling into the water would be to maintain that there is some other possible world like w_2 that is at least as similar to w as is w_1.[6] And to avoid making the position of the boat causally dependent on George's falling in the river, one would also have to maintain that worlds such as w_1 are at least as similar to w as are worlds like w_2. This is the view to which my reformulation of Lewis's account is committed in general, and I think that in this case Lewis would accept it, too. He writes, "[W]e should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future. That is not to say, however, that the immediate past depends on the present in any very definite way" (1979, p. 40). Counterfactual theorists have a choice among implausibilities. They can say that w_3 is more similar to w than is w_1 or they can say that there are ties between worlds like w_1 and w_2. The fact that my account is committed to the second alternative is no argument for favoring Lewis's formulation. I personally do not think that causation is counterfactual dependence. I am only arguing that my reformulation of a counterfactual theory is at least as plausible as Lewis's account.

The independence condition (**I**) implies that the causes of an event are not all causally connected to one another. If one assumes that counterfactual dependence implies causal connection, that is, that distinct events that are not causally connected are not counterfactually dependent on one another, it follows that not all the causes of an event will be counterfactually dependent on one another. The independence condition plus the assumption that counterfactual dependence implies causal connection also implies that effects of a common cause will each have their own causes that are not counterfactually independent on one another. Given the view that there will be ties between non-*b* possible worlds missing one or another of *b*'s independent causes, causes will not be counterfactually dependent on their effects and effects of a common cause will not be counterfactually depend-

[6] I am indebted here to Horacio Arló-Costa.

ent on one another. For a more rigorous presentation of this argument, see the proof of theorem 6.1, p. 135.

One can go on to deduce the sufficient condition **L** states (that counterfactual dependence implies causation) from the same premises plus the connection principle. If *b* counterfactually depends on *a*, then *a* and *b* must be causally connected. Since causes are not counterfactually dependent on their effects and effects of a common cause are not counterfactually dependent on one another, the only way that *a* and *b* can be causally connected is for *a* to be a cause of *b*.

It thus seems that one might defend the simplified version of Lewis's theory, **L**, with a different explanation for why individual causes are not causally dependent on their effects and why effects of a common cause are not causally dependent on one another. Given the independence condition, one need not rely on miracles, and one can ground the asymmetry of causation in the multiplicity of independent causes. There is however little reason to develop such a counterfactual theory as an *alternative* to the independence theory of causal priority, **CP**, because someone accepting the independence condition, the connection principle, and transitivity, is already committed to **CP**. By accepting **CP**, one can avoid the difficulties of analyzing counterfactuals and of defending an account of similarity among possible worlds.

Qualms about the independence condition are no reason to prefer Lewis's theory, because Lewis assumes that there is an asymmetry of overdetermination, which is closely related to the independence condition (see theorem 6.3, p. 136). Moreover, if one clings to Lewis's semantics and interprets miracles as interventions, then the claim that effects are not counterfactually dependent on their causes entails the strong independence condition (which in turn entails the independence condition **I**). Accepting Lewis's account is tantamount to assuming that there will always be something that prevents the effect and leaves the causes alone – that is, that there will always be something that is causally independent of the given causes.

6.4 Counterfactuals and Predictions

People sometimes reason counterfactually from present to past. Lewis maintains that in doing so, they interpret counterfactuals in a nonstandard way (1979, p. 34; see also Bennett 1984). Is this plausible? The alternative discussed above permits one to deny the counterfactual dependence of individual causes on their effects without invoking well-placed miracles, and it permits counterfactuals of the same kind forward and backward in time (see also Bennett 1984; Goggans 1992).

This seems to me a virtue, even though Lewis and others have argued that one cannot combine backward and forward counterfactuals lest one wind up maintaining inconsistently both "If he had jumped out of the window, he would have broken his neck" and "If he had jumped out of the window, he wouldn't have broken his neck" (because he would never have jumped without first installing a safety net). I am convinced by Jonathan Bennett's response, "Different standards for closeness to the actual world are *arbitrarily* associated with different temporal directions" (1984, p. 71). One can get the same inconsistencies, the same impossibility of finding a single way of determining closest or most similar possible worlds with respect to two forward conditionals. No single interpretation makes both counterfactuals true, because in endorsing one, we are holding fixed propositions that are inconsistent with those we hold fixed in endorsing the other. As pointed out in the above discussion of George jumping from the bridge, Lewis's own semantics for counterfactuals already requires some conditionals going backwards in time. Not only does the reformulation of Lewis's theory sketched in the previous section permit one to avoid ad hoc reliance on miracles, it also permits a unified treatment of counterfactuals both forward and backward in time.

There is another reason to favor the above account over Lewis's rules for comparing possible worlds. Counterfactual reasoning should permit one to work out the implications of counterfactual suppositions, so as to be prepared in case what one supposes actually happens. A child asks herself, "What will happen if I push that button?" and finds out by pushing the button and seeing what does happen. That can be a dangerous way to get an answer. It is handy to be able instead to pose counterfactual questions and to rely on their answers to predict what will happen. An analysis of counterfactuals ought to tie their truth closely to the truth of predictions concerning what will happen if . . . . A counterfactual of the form, "If I were to push the button, the alarm would go off," ought to license one to predict that the alarm will go off if one in fact pushes the button. Such a prediction may go astray, because of independent changes in other causes of the alarm's sounding, but such predictions must nevertheless be justifiable. Counterfactuals should satisfy a prediction condition.

Lewis's account of similarity among possible worlds implies that knowledge of counterfactuals will justify predictions about the results of *interventions*, which one can model as miracles. Consider the counterfactual: "If I were to push the button, the alarm would go off." On Lewis's account, in the closest possible worlds in which I push the button the other causal factors upon which the alarm sounding depends are unchanged, and my pushing the button is due to a minor difference between the laws of the actual world and the laws of this possible world. The alarm sounds in this possible world if and only if one can predict that it will sound in the real world when one *intervenes* and pushes the button. By definition, a button push that is brought about by an intervention is independent of changes in the values of any other variables, so the counterfactual is true if and only if the best prediction of what will actually happen if I push button is that the alarm sounds.

So far so good. But I would argue for a stronger prediction condition, which does not apply only in cases of intervention:

> **P** (*Prediction condition*) The knowledge that *b* would occur if *a* were to occur and that an event of kind **a** occurs taken by itself justifies the prediction that an event of kind **b** will occur.

By "taken by itself" I mean that one has no knowledge concerning the circumstances – that is, concerning whether other causes occur or other causal conditions obtain. Events of kind **a** instantiate the same causally relevant property that *a* does. The prediction condition says that knowledge of counterfactual dependence among tokens justifies the prediction that if a cause of the same kind occurs, then (ceteris paribus) an effect of the same kind will occur. **P** bears a vague resemblance to modus ponens: From "if *a* were to occur then *b* would occur" and "an event of kind **a** occurs" infer (fallibly) "an event of kind **b** occurs."

Counterfactuals should guide our thinking about nonexperimental as well as experimental situations. Regardless of whether one is concerned with the results of interventions or with the results of mere happenings, knowledge of the value of x should provide more "guidance" concerning the value of y when one knows that the value of y counterfactually depends on the value of x than when one does not know this. Knowledge of counterfactual dependencies should be reflected in our expectations about what will happen.

Consider, for example, the apparatus in figure 6.2, in which a salt solution flows through the pipes and mixing chambers. With constant flows, the salt concentration in chamber 5, x_5, is the average of the concentrations in chambers 3 and 4: $x_5 = .5(x_3 + x_4)$. This equation expresses a causal generalization: x_3 and x_4 influence x_5 in just this way. Paralleling the causal dependencies are apparently asymmetrical counterfactual dependencies. For example, the value of x_5 counterfactually depends on the value of x_4 and not vice versa. **P**, the prediction condition, says that knowing the counterfactual dependency of the value of x_5 on the value of x_4 justifies predicting that when one measures an increase in the concentration of basin 4, the concentration of basin 5 increases by half as much. Measuring an increased concentration in basin 5 does not, in contrast, justify any prediction concerning the concentration in basin 3 or the concentration in basin 4. It justifies only a prediction that one or both of these concentrations changes or that something changes in the mechanism carrying salt solution to basin 5. According to the prediction condition, counterfactual depend

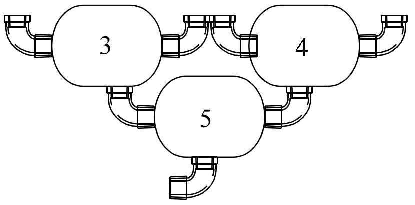


Figure 6.2: A simple causal structure

ence makes this predictive difference.

Lewis's account does not satisfy the prediction condition. Consider the following four counterfactuals:

1. If *a* had occurred "miraculously" (such as via an intervention), then *b* would have occurred.
2. If *a* had occurred as a consequence of *d*, then *f* would have occurred.
3. If *a* had occurred, then *b* would have occurred.
4. If *a* had occurred, then *f* would have occurred.

According to Lewis's theory, if (1) is true, then (3) is true and (4) is false, regardless of whether or not (2) is true. One knows that (1) is true if and only if one knows that (3) is true. Suppose we know that (1) and (2) are both true and that some token of type **a** occurs. One cannot justifiably predict that a token of kind **b** will occur. Whether a token of kind **b** or of kind **f** occurs depends on what causes the token of kind **a**. The prediction condition says that if we know that (3), then a prediction that an event of kind **b** will occur will be justified (unless we have extra information concerning the state of other causal factors). Since the prediction is not justified and we do not have such extra information, we do not know that (3). By assumption we know that (1). Hence Lewis is mistaken to identify (1) and (3).

For example, engineers checking the design of a nuclear power plant may ask, "What would happen if that steam pipe were to burst?" They want their answer to match what one will observe in the event that the pipe actually does burst, though they hope never to make the observation. According to Lewis, they should consider a possible world exactly like ours until near the time of the pipe bursting, at which point some small miracle occurs, and the world evolves according to laws of nature like ours. In such a world, let us suppose that the reactor shuts down promptly.

The bursting of the pipe may have different consequences when it bursts because of an earthquake. The engineers will not and should not necessarily assume that the pipe burst because of a small miracle immediately preceding the bursting and they will not and should not predict that the consequences of the pipe's bursting will be that the reactor shuts down promptly. Knowledge that the pipe burst does not by itself justify *any* prediction about whether the reactor will shut down. The engineers need to do some backtracking and to say, "If the pipe were to burst, then either it was faulty, or a girder fell on it, or there was an earthquake, or there was sabotage, or the pressure became too great. The consequences of the bursting depend on which of these holds." *Responsible engineers must do such backtracking when the consequences of the pipe's bursting depend on what caused it to burst*. If the pipe burst because the pressure was too great, and the pressure was too great because the reactor was going out of control, then the consequences of the pipe bursting may be different than if it were caused by corrosion, a faulty weld, or a terrorist's bomb. In order to consider how the world would differ in the future in consequence of the bursting, the engineers must also think about how the world must have differed in order for the bursting to have occurred. Lewis's suggestion that the most similar possible worlds involve some small miracle just before the pipe bursts rules out the above reasoning, and it is for this reason mistaken. The counterfactual, "If the pipe were to burst, then the reactor would shut down safely" is *false*. No prediction is justified concerning the consequences of the bursting for whether the reactor shuts down until one specifies what caused the bursting.

The alternative view of counterfactuals sketched above (and developed in more detail in §6.2*) permits one to "hold fixed" laws of nature, not only in considering the consequences of *a*'s occurring, but also in considering what would have caused *a* to occur. It may not matter what caused some event to occur or not to occur, and then there is no harm in supposing it occurred in one specific way, such as by a miracle. But it may matter, and one will need to explore how the bursting could have followed lawfully from its causes (Bennett 1984, pp. 72–4). In considering what would happen if *a* occurred, one possibility is that *a* happened inexplicably, as if by intervention or miracle, but it would be irresponsible to suppose that that is the *only* way *a* might happen. If one accepts the prediction condition, then one should deny that worlds that differ in laws are more similar to the actual world than are worlds that differ in causal antecedents.

The quantitative example I introduced above underlines this point. In illustrating the asymmetry of counterfactual dependence and the content of the prediction condition, I assumed that the concentrations in basins 3 and 4 were independent of one another. But suppose x_3 and x_4 are not independent. Suppose that the apparatus is as shown in figure 6.3. The salt concentration in both chambers 3 and 4 depends on the salt concentration in chamber 1. Now one cannot justifiably predict what the concentration in basin 5 will be if the salt concentration in chamber 4 increases by z^*. If one insists on last-minute miracles, one will predict that the concentration in chamber 3 remains unchanged and the concentration in chamber 5 increases by $.5z^*$. But this prediction is not justified. When one measures an increase

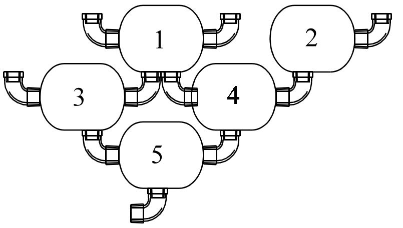


Figure 6.3: A structure with multiple connection

in the concentration in basin 4, the concentration in basin 3 need not remain unchanged, even if there are no independent changes. Nor need the concentration in basin 5 increase by .5*z**. These generalizations would be true if x_3 and x_4 were independent of one another, but they are not.

The consequences of the increase of the concentration in chamber 4 depend on its causes. If the increase in the concentration in basin 4 were due entirely to an increase in the concentration in basin 2, then the value of x_5 would increase by .5z*. If the doubling were due to an increase in the concentration in basin 1, then the value of x_5 would increase by more than .5*z**. The concentration in x_4 might also result from an intervention such as the addition of salt through a hatch at the top of basin 4. A possible world in which an altered value of x_4 results from an intervention is no more similar to the actual world than is a possible world with a change in the values of x_1 or x_2 or both. The causes of the value of x_4 matter to its effects: in other words, x_4 does not screen off its causes from its effects. (An event *a* screens off its causes from a direct effect *b* if and only if *a* is independent of all other proximate causes of *b*.) Insisting on always postulating miracles at the last possible moment leads to a violation of the prediction condition. It is not the right way to prepare us to deal with actual happenings.

6.5 Critique of Lewis's Account of Similarity Among Possible Worlds

Should one conclude that Lewis is mistaken concerning similarities among possible worlds? In defense of Lewis, one can point out that only the state of the world at the moment when the pipe bursts or when there is a change in the salt concentration in basin 4 matters. Rather than inquiries into what caused the bursting or the change in salt concentration, one needs a correct specification of the values of all relevant *contemporary* variables. Once one specifies the value of x_3, there's no problem inferring the value of x_5 from counterfactual suppositions concerning the value of x_4. (Though with all the causes but one specified, there is equally little problem with the reverse inference.) Alternatively, a defender of Lewis's semantics can point out that one can consider counterfactuals with more complicated antecedents such as "If the pipe had been faulty and then had burst," or "If the pressure had become too great and the pipe had burst."

These maneuvers do not rescue Lewis's account. One is interested in the consequences of the pipe's bursting given the values of other relevant variables that one may actually encounter. And to determine what those values are, one needs to consider what might have caused the pipe's bursting, so that one can determine whether the other relevant variables depend on these causes. Similarly, one needs to backtrack to decide whether one needs to consider counterfactuals with more complicated antecedents and, if so, which ones one should consider. In determining the implications of a counterfactual supposition such as "What if this pipe burst?" one must backtrack.

A defender of Lewis's semantics could concede that backtracking is needed, yet maintain that backtracking plays only an epistemological role in the evaluation of counterfactuals. One backtracks to determine which counterfactual question to ask, not to give the answer. Backtracking helps one to transform counterfactual suppositions into counterfactual claims, not to determine whether counterfactual claims are true. Backtracking may be prevalent and important, but it has no role when one is stating the truth conditions of counterfactuals.

This response will not do. Suppose one tries to assess the counterfactual:

(C) If the value of x_4 were to increase by z^*, then the value of x_5 would increase by $.5z^*$.

Consider figure 6.4. The actual situation is depicted in figure 6.4a. $x_1 = x^*_1$, $x_2 = x^*_2$, $x_3 = x^*_3$, $x_4 = x^*_4$, $x_5 = x^*_5$ and the causal relations are as shown. Consider then the possible situations depicted in figure 6.4b, c, and d. In each of these, x_4, the concentration in basin 4, increases to $x^*_4 + z^*$. But the source of the increased concentration differs. In w_1 depicted in figure 6.4b, the increase is due to some difference between the laws of w_1 and the actual world (or to some unspecified mechanism or intervention), which leaves the values of x_1 and x_2 and x_3 and the causal structure relating x_3 and x_4 to x_5 unaffected. In w_2, shown in figure 6.4c, the increase in x_4 occurs because of an increase in x_1. In w_3, shown in figure 6.4d, the increase in x_4 is due to an increase in x_2. In w_1 and w_3 the value of x_5 is larger than it is in the actual

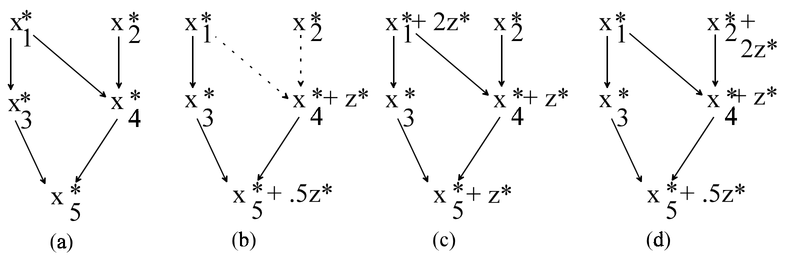


Figure 6.4: Which is closest?

world by .5z*, while in w_2 the value of x_5 increases by z*.

If we set aside questions suggested by Lewis's talk of "orderly transitions" (1979, p. 40), Lewis's semantics says that w_1 is closer to the real world than w_2 or w_3, and the counterfactual **C** is true. w_1 is closer to the actual world because the period of "perfect match" between w_1 and the actual world is longer than the period of perfect match between the actual world and either w_2 or w_3. This is a weak reason. Why should a few seconds of additional perfect match be decisive? Lewis answers in effect that unless a few seconds more match are decisive, one cannot account for our judgments of counterfactual and causal dependence. The alternative theory presented here undercuts this reason, because it accounts for our judgments concerning counterfactual and causal dependence without requiring that w_1 be more similar to the real world than is w_2 or w_3. All that remains is the dubious intuitive argument that any increase in the period of perfect match makes for greater similarity.

According to my account w_2 and w_3 are at least as similar to the actual situation as is w_1, and the counterfactual **C** is false. One needs to know how the increase in the salinity of basin 4 occurred before one can make any predictions or any true counterfactual claims concerning how much the concentration in basin 5 would increase. The falsity of the counterfactual **C** results from the "double-connection" between x_1 and x_5 – one connection via x_3 and one via x_4. If x_1 were not a cause of x_3 as well as x_4 – if x_3 and x_4 were causally independent of one another – there would be no need to bring x_1 or x_2 into the picture.

This account denies that non-*b* possible worlds involving miracles immediately before *b* are more similar to the actual world than non-*b* possible worlds without miracles right there. This account does not maintain that the most similar possible worlds will be free of miracles. It is plausible to maintain, as Lewis does, that possible worlds with completely different histories are very unlike the actual world. When there are no multiple connections between a cause and one of its effects, or, equivalently, when the causes of an event *b* do not matter to its effects, then one is free to suppose that *b* failed to occur by a miracle. There is no requirement that one keep backtracking endlessly. One stops backtracking when it no longer matters to the consequences of a supposition how the changes in causal ancestors came about, or when one comes to an event whose direct causes are all causally independent of one another. The context, one's purposes, and the causal relations enable one to isolate a "system" of interest in which miracles should be avoided.

6.6 Refutation of L and Defense of a Restricted Version

In §6.3 I showed how a view of comparative overall similarity that did not favor miracles, coupled with other plausible conditions, implies a simplified formulation of Lewis's sufficient condition for causation and permits one to deny that individual causes counterfactually depend on individual effects or that effects of a common cause counterfactually depend on one another. In the last sections I argued for this alternative view and against Lewis's account of similarity on the grounds that Lewis's account does not permit backtracking and that it implies the truth of counterfactuals that do not justify predictions. If one then accepts my alternative account, one must give up any hope of providing a noncircular counterfactual theory of causality, because similarity among possible worlds and the truth of counterfactuals would depend on explicitly causal facts.

One might, however, hope to defend a theory of the relations between causation and counterfactuals similar to the one sketched in Herbert Simon and Nicholas Rescher's essay, "Cause and Counterfactual" (1966). They link the asymmetry of causation to an asymmetry of counterfactual determinacy. Particular effects counterfactually depend on each of their causes, while particular causes do not counterfactually depend on any of their effects.

This hope cannot be sustained, because the necessary condition that **L** states is false. Causal dependence does not imply counterfactual dependence. The concentration in basin 5 causally depends on the concentration in basin 4, but the concentration in basin 5 does not counterfactually depend on the concentration in basin 4. Simon and Rescher's claim about the connections between causation and counterfactuals and Lewis's theory are both mistaken.

This difficulty does not derive from my suppression of Lewis's distinction between causal dependence and causation. Lewis's theory permits x_4 to be a cause of x_5 without any counterfactual dependence of x_5 on x_4. All that's needed is a chain of counterfactual dependence. But there is no chain of counterfactual dependence here. The salt concentration at the top of the pipe between basins 4 and 5 counterfactually depends on a change in concentration in basin 4, and the concentration at the bottom of the pipe counterfactually depends on the concentration at the top of the pipe. But the concentration in basin 5 does not counterfactually depend on the concentration at the bottom of the pipe. This is a case of causation without any chain of counterfactual dependence.

One might attempt to defend **L** as follows: "Regardless of how the change in salt concentration in basin 4 comes about, if the concentration in basin 4 were to change, then so would the concentration in basin 5. So there is after all no difficulty with **L**." This defense is unsatisfactory, and not only because Lewis also wants his account to apply to quantitative causal dependence (1973a, p. 166). Suppose in figure 6.3 that device number 1 is not a basin holding a salt solution, but some mechanism that generates exactly neutralizing quantities of an acid and an alkali, which flow through the pipes to basins 4 and 3, respectively. If the change in the acidity in basin 4 is due to a greater acid output from device 1, then there will be no change in the acidity of 5, because the greater acid output from 1 transmitted to 4 is neutralized by an increase in alkali output transmitted to 3. It would thus be false to say, "If the acidity of 4 were greater, then the acidity of 5 would be greater." If, for example, one had a switch which enabled one to affect the combined acid-base output rate of device 1, then when one moved the switch, the acidity of 4 would change without any change in the acidity of 5. One can give a similar qualitative example. It could be the case that whether the solution in 5 is acid at all does not depend counterfactually on whether there is any acid in 4. Yet the acidity of 5 causally depends on the acidity of 4. Causal dependence does not imply counterfactual dependence.

If one insists that causal dependence *must* be reflected in counterfactual dependence, one must take counterfactual suppositions, such as "if the acidity of 4 had been different" as suppositions that the differences came about via miracles. In that case, as already explained, one will be in the position of saying, "If *a* were to occur, *b* would occur, but one cannot predict whether events of kind **b** will occur when events of kind **a** occur." Second, the grounds for accepting Lewis's view of similarity among possible worlds would be one's knowledge of causal relations and one's desire to equate causal and counterfactual dependence. One would have tacitly abandoned the attempt to give a counterfactual theory of causation and in its place one would be offering a causal theory of counterfactuals – the theory of similarity among possible worlds would be grounded in an account of causation. **L** is false.

Let us say that there is a "multiple connection" between *a* and *b* if some cause *d* of *a* is or in the absence of *a* would be connected to *b* by a path that does not go through *a* (as x_1 is connected to x_5 via both x_3 and x_4). There will be a multiple connection between *a* and *b* if and only if controlling for events of kind **a** in these circumstances does not screen off *b*'s from the causes of *a*'s. If there is a multiple connection between *a* and *b*, then *b* will not counterfactually depend on *a* just as *a* does not counterfactually depend on *b*. One will have a case of causation without any chain of counterfactual dependence, and there will be no asymmetry of counterfactual dependence.

L can, however, be defended as an approximate truth, and a restricted form of **L** can be proven. When there are no multiple connections, **L** is true. Since situations involving multiple strong connections are rare, **L** is a good approximation. Given the connection principle and the revised account of similarity among possible worlds that is implicit in the discussion above, it can be proven that if there are no multiple connections between cause and effect, then causation implies counterfactual dependence. This claim is formulated more rigorously as theorem 6.4. Lewis's theory, restricted to circumstances in which there are no multiple connections, follows from a revised account of similarity among possible worlds, the claim that if distinct events are counterfactually dependent then they are causally connected, and the independence theory of causal priority presented in chapter 4.

6.7 What Does the Counterfactual Theory of Causation Teach Us?

The view of counterfactuals sketched in this chapter is not equivalent to:

> *A mistaken view*: If **P** were the case then **Q** would be the case if and only if **Q** is deducible from a nonredundant conjunction of statements including **P**, laws of nature, and specifications of the circumstances.

Consider the relationship between the salt concentrations in basins 1, 2, and 4 in figure 6.3, and suppose that the concentrations in basins 1 and 2 are causally independent of one another. If concentration in basin 2 is part of the "circumstances," one can deduce the concentration in basin 4 from the concentration in basin 1 or vice versa. If the concentration in basin 2 is not part of the circumstances, one can make no deductive inference in either direction. Deduction from laws and circumstances implies a *symmetrical* relationship between x_1 and x_4, but the counterfactual dependence here is asymmetrical, since there is (by assumption) no multiple connection. The concentration in basin 4 is counterfactually dependent on the concentration in basin 1, while the concentration in basin 1 is not counterfactually dependent on the concentration in basin 4. Where does the asymmetry come from?

The asymmetry arises from the fact that the values of x_1 and x_2 are causally connected to the value of x_4, but independent of one another. One has an "intransitive triplet" (Pearl and Verma 1994, p. 804). So when one supposes that x_1 has some other value, one "holds fixed" x_2, but when one supposes that x_4 has some other value, one does not hold fixed x_2, because the different value of x_4 may be due to a difference in the value of x_2. This is not a human quirk. If one measures different values of x_1 and predicts the value of x_4 on the assumption that the value of x_2 remains fixed, one will not necessarily be right, because the value of x_2 may vary independently. But if x_1 and x_2 are independent and one knows nothing about how x_2 may have varied, the calculated values of x_4 will be the best prediction of the actual values. This *fact* justifies making the prediction. If one measures different values of x_4 and calculates values of x_1 on the assumption that x_2 remains fixed, the calculated values will not be the best predictions of the actual values of x_1. On the contrary, the variations in the measured values of x_4 constitute evidence that the value of x_2 has changed. Thus statisticians prove that independence of error terms – of the other causes – is a necessary condition for unbiased estimation (Festa 1993, p. 39).

When there are no multiple connections, a causal intermediary *a* screens off its causes from its effects. It does not matter what caused *a* to occur. The other factors that contribute directly to the effects of *a* are causally independent of *a*. So one can hold them fixed, and the effects of *a* are individually counterfactually dependent on *a*. When there are multiple connections, as in the example of the basins, some of the causes of the given effect are not causally independent of one another (as, for example, the values of x_3 and x_4), and it is no longer the case that one should "hold fixed" the value of one when one supposes that the value of the other changes. It is the independence of causes that permits the counterfactual dependence of individual effects on individual causes and defeats the counterfactual dependence of individual causes on individual effects. But when there are multiple connections, not all the causes are independent, and individual effects do not counterfactually depend on individual causes.

These links between independence and counterfactual dependence establish a restricted version of **L**. **L** is thus not a tenable alternative account of causal priority. It is not tenable for the reasons already spelled out in this chapter. It is not an alternative, because it more or less presupposes the independence condition and, in deterministic single-connection circumstances, it follows from the independence theory, the claim that causal connection is a necessary condition for counterfactual dependence among distinct events, and a view of similarity among possible worlds that is at least as plausible as Lewis's.

Even if unacceptable as a theory of causation, a counterfactual theory can still tell us some things worth knowing about causation. Although there are cases in which *a* is a cause of *b*, even though *b* does not depend counterfactually on *a*, such cases are infrequent unless the causal connection between *a* and *b* is remote. It is seldom the case that there are multiple connections in which more than one of the connections is strong enough to worry about. Consequently, it is a good first approximation to say that there is an asymmetry of counterfactual dependence. It is worth noting the connections between causation and counterfactual dependence, even if one cannot defend a counterfactual theory of causation.