CAUSATION

Chapter 5

Humean Reductionism - Counterfactual Approaches

A second important reductionist approach attempts to analyze causation in terms of counterfactuals. Such approaches come in different forms, and can be arrived at via different routes. One way of arriving at a counterfactual account is by analyzing causation in terms of necessary and/or sufficient conditions, but then interpreting the latter, not in terms of nomologically necessary and nomologically sufficient conditions, but in terms of subjunctive conditionals. Thus, one can say that *c* is necessary in the circumstances for *e* if, and only if, had *c* not occurred, *e* would not have occurred, and that *c* is sufficient in the circumstances for *e* if, and only if, had *e* not occurred, *c* would not have occurred.

John Mackie took this tack in developing a more sophisticated analysis of causation in terms of necessary and sufficient conditions. Thus, after defining an INUS condition of an event as an insufficient but necessary part of a condition which is itself unnecessary but exclusively sufficient for the event, and then arguing that *c*'s being a cause of *e* can then be analyzed as *c*'s being at least an INUS condition of *e*, Mackie asked how necessary and sufficient conditions should be understood. For general causal statements, Mackie favored a nomological account, but for singular causal statements, he argued for an analysis in terms of subjunctive conditionals (1973, p. 48).

The most fully worked out counterfactual approach, however, is that of David Lewis (1973).[1] His original, basic strategy involved analyzing causation in terms of a narrower notion of causal dependence, and then analyzing causal dependence counterfactually: (1) an event *c* causes an event *e* if and only if there is a chain of causally dependent events linking *e* with *c*; (2) an event *g* is causally dependent upon an event *f* if and only if, had *f* not occurred, *g* would not have occurred.

Causes, so construed, need not be necessary for their effects, since counterfactual dependence, and hence causal dependence, are not necessarily transitive. Nevertheless, Lewis's approach is very closely related to necessary condition analyses of causation, since the more basic relation of causal dependence is a matter of one event's being counterfactually necessary in the circumstances for another event.

[1] For later discussion, and some revisions, see David Lewis (1979 and 1986).

5.1 Some Important Objections to David Lewis's Counterfactual Analysis of Causation

How satisfactory are analyses of causation in terms of counterfactuals? One objection to Lewis's approach is that it is formulated in terms of events, and it then becomes a delicate matter to set out an account of the individuation of events that will not generate unwelcome consequences concerning causal relations (Lewis, 1986b, and Bennett, 1998). A much better approach, it would seem, would be to view the ***basic*** causal relata as states of affairs - or as events in Jaegwon Kim's sense (1971, 1973a) - and thus to regard the ***basic*** singular causal statements as those that explicitly specify the causally relevant factors, and that do not incorporate causally extraneous information. For not only does this seem metaphysically more perspicuous, it also enables one to avoid getting one's account of causation entangled in the problem of the individuation of events.

Of course, we certainly make causal statements that provide no information at all about what properties and/or relations enter into the causally relevant states of affairs - such as "Mary's remark caused an interesting occurrence." But it would seem to be a relatively straightforward matter to analyze such event-statements in terms of metaphysically more basic statements concerning causal relations between states of affairs.

A second objection - originally advanced by Jaegwon Kim in his article, "Causes and Counterfactuals" (1973b) - focuses upon the fact that there are a number of counterfactuals that have nothing to do with causation. If, for example, John and Mary are married at time *t*, it is true that if John had not existed at time *t*, then Mary would not have been married at time *t*. But John's existing at time *t* is not a cause of the simultaneous state of affairs that is Mary's being married at time t.

How might this objection be handled? There has been relatively little discussion of this problem, but it would seem that one will have to draw a line between counterfactuals whose truth depends upon laws of nature, and those whose truth does not so depend. Exactly how this is to be done, given the type of approach to counterfactuals that must be employed here, is not entirely clear.

A third objection is that some counterfactuals are based upon non-causal laws. Thus, for example, counterfactuals such as "If *A* had not exerted force *F* upon object *B*, then *B* would not have exerted a force *G* upon *A*" will be true in Newtonian worlds by virtue of Newton's Third Law of Motion. On Lewis's counterfactual analysis, it follows that *A*'s exerting force *F* on *B* causes *B*'s exerting force *G* on *A*, and vice versa - which is surely wrong.

A fourth objection involves overdetermination, or redundant causation, where two events, *C* and *D*, are followed by an event *E*, and where each of *C* and *D* would have been causally sufficient, on its own to produce *E*. If it is true that *C* causes *E* and that *D* causes *E*, then one has a counterexample to Lewis's counterfactual analysis. Lewis contends that we are uncertain what to say here. Do *C* and *D* each cause *E*, or do they jointly cause *E*? But is Lewis right about this? If, for example, Lewis (1973, p. 73) were right in holding that "a contingent generalization is a *law of nature* if and only if it appears as a theorem (or axiom) in each of the deductive systems that achieves a best combination of simplicity and strength," then it would seem that it would have to be the case that *C* causes *E* and *D* causes *E*, since more complicated generalizations are needed if one is to say instead that it is only the combination of *C* together with *D* that causes *E*. Similarly, if simplicity is, instead, epistemologically relevant, will not the conclusion be the same? So overdetermination certainly seems to be a problem.

A fifth objection involves cases of preemption - where, once again, one has causation without causal dependence. Until recently, the discussion of preemption had focused on cases where one causal process preempts another by blocking the occurrence of some state of affairs in the other process, and a variety of closely related ways of attempting to handle this type of preemption have been advanced, involving such notions as fragility of events, quasi-dependence, continuous processes, minimal-counterfactual sufficiency, and minimal-dependence sets (Lewis, 1986c; Menzies, 1989; McDermott, 1995; Ramachandran, 1997). But none of these approaches can handle the case of trumping preemption, advanced by Jonathan Schaffer, in his article "Trumping Preemption" (2000).

David Lewis's own reaction to the problem posed by trumping preemption has been to replace his previous counterfactual accounts by a new, 'causation as influence', account:

> Where *C* and *E* are distinct actual events, let us say that *C* *influences E* if and only if there is a substantial range C_1, C_2, . . . of different not-too-distant alterations of *C* (including the actual alteration of *C*) and there is a range E_1, *E*2 . . . of alterations of *E*, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so on. (2000, p. 190)

But this account does not really provide an answer to the trumping preemption objection. For suppose that, contrary to what is required for Lewis's idea of causation as influence to be applicable, there is *not* a substantial range C_1, C_2 . . . of different not-too-distant alterations of *C*: there is only *C*, or its absence. Suppose further that there is a substantial range of alterations of *D* - D_1, D_2, and so on - and where, in the absence of *C*, *D* will give rise to *E*, D_1 to E_1, D_2 to E_2, and so, and where *E*, E_1, E_2, etc. are all distinct. Suppose, finally, that if *C* accompanies any of *D*, D_1, D_2, etc., then it is always *E* that comes about. Then surely the simplest hypothesis will involve laws according to which *C* preempts all of *D*, D_1, D_2, etc. Consequently, there are cases of trumping preemption that Lewis's revised account cannot handle.

5.2 The Fundamental Objection to Counterfactual Analyses of Causation

The preceding objections are all, I think, important ones, and I am inclined to think that at least some of them are likely to constitute decisive objections to the counterfactual approach to causation. However, as the ongoing discussions of, for example, preemption show, to show that any of these objections provides a refutation of all counterfactual analyses of causation calls for considerable work. My goal in this paper, accordingly, is to pursue a different line of attack, and one that can, I think, be shown to be decisive.

If causation is to be analyzed counterfactually, one needs to show that there is a satisfactory account of counterfactuals that is compatible with such an analysis. Many accounts of counterfactuals are clearly not available, since they incorporate causal concepts. This is so, for example, of the analysis of counterfactuals advanced by Frank Jackson in his article, “A Causal Theory of Counterfactuals” (1977), and it is also the case for the somewhat complex account advanced by Igal Kvart in his book, *A Theory of Counterfactuals* (1986). The question, then, is whether it is possible to provide a satisfactory analysis of counterfactuals without employing causal notions. If it can be shown that no such account is available, then a counterfactual analysis of causation does not even get started.

5.2.1 The Stalnaker-Lewis Approach to Counterfactuals

There is, of course, an obvious candidate to play this role: a Stalnaker-Lewis-style analysis of counterfactuals. After all, many philosophers today regard this approach as the standard account of the truth conditions of counterfactuals. So why is there a problem? The answer, as I shall try to show, is that there are decisive objections to a Stalnaker-Lewis approach to counterfactuals.

5.2.2 The Nature of the Account

This general approach to counterfactuals–which appeals to similarity relations between possible worlds–was first set out by Robert Stalnaker in his 1968 article, "A Theory of Conditionals,"[2] and then a modified, and in some ways more satisfactory version of it was advanced by David Lewis in his 1973 book, *Counterfactuals*.[3]

[2] Robert C. Stalnaker, "A Theory of Conditionals," in Nicholas Rescher (ed.), *Studies in Logical Theory* (Oxford: Blackwell, 1968).

[3] David Lewis, *Counterfactuals*, (Cambridge, Massachusetts: Harvard University Press, 1973).

David Lewis's detailed exposition, in his book *Counterfactuals*, of this type of account of counterfactuals elicited some very important criticisms of the whole approach, including ones advanced in early reviews by Jonathan Bennett[4] and Kit Fine.[5] But Lewis, in "Counterfactual Dependence and Time's Arrow,"[6] and then in his "*Postscripts to* 'Counterfactual Dependence and Time's Arrow',"[7] attempted to show that, given, for example, the right sort of account of the factors that enter into judgments of similarity between possible worlds in the case of counterfactuals, one could escape the most crucial objections that had been advanced.

One of my goals in this paper is to show that while the answer that Lewis offered to the most fundamental objection advanced by Jonathan Bennett and Kit Fine is successful in blocking the objection as they stated it, it fails when confronted with a reformulation of that objection. But I shall also be arguing that the approach is exposed to other, very strong objections.

5.2.2.1 Robert Stalnaker's Proposal in "A Theory of Conditionals"

In his article, "A Theory of Conditionals," Robert Stalnaker offered the following, "selection-function" account of counterfactuals:

" . . . our semantical apparatus includes a *selection function*, *f*, which takes a proposition and a possible world as arguments and a possible world as its value. The *s*-function selects, for each antecedent *A*, a particular possible world in which *A* is true. The *assertion* which the conditional makes, then, is that the consequent is true in the world selected. A conditional is true in the actual world when its consequent is true in the selected world."[8]

But exactly what idea is the selection function supposed to be capturing? Stalnaker's answer is that the informal truth conditions that he proposed earlier in the article "required that the world selected *differ minimally* from the actual world." Thus, "the selection is based on an ordering of possible worlds with respect to their resemblance to the actual world."

[4] Jonathan Bennett, "Counterfactuals and Possible Worlds," *Canadian Journal of Philosophy*, 4 (1974), 381-402.

[5] Kit Fine, "Critical Notice - *Counterfactuals*," *Mind*, 84 (1975), 451-8.

[6] David Lewis, "Counterfactual Dependence and Time's Arrow," *Noûs*, 13 (1979), 455-76.

[7] David Lewis, "*Postscripts to* 'Counterfactual Dependence and Time's Arrow'," *Philosophical Papers*, Volume II (New York: Oxford University Press, 1986).

[8] Stalnaker, "A Theory of Conditionals," in Sosa, p. 170.

What reason did Stalnaker offer for embracing this sort of account? The answer is that his acceptance does not seem to rest upon much beyond some criticisms that he advances against a particular type of alternative - namely, one that appeals to logical or causal connections between the antecedent and the consequent of a counterfactual. But such criticism is very narrowly directed. It does not, for example, give one any reason for rejecting causal approaches in general, since the criticism in question does not tell against approaches according to which "If p were the case, then q would be the case" can be true even though p does not itself have any causal or logical connection to q, because, for example, there is some r that is compatible with p, and which is true, and which is causally related to q.

5.2.2.2 David Lewis's Account

In his book, *Counterfactuals*, David Lewis offered the following, "ordering relation" account of counterfactuals:

> $\phi \;\Box\!\!\rightarrow \psi$ is true at a world *i* (according to a system of spheres \$) if and only if either
>
> (1) no ϕ-world belongs to any sphere *S* in $\$_i$, or
>
> (2) some sphere *S* in $\$_i$ does contain at least one ϕ-world, and $\phi \supset \psi$ holds at every world in *S*.[9]

What is the "system of spheres"? The basic idea is that for any possible world, all other possible worlds can be placed on spheres that are centered on the world in question, with the size of a given sphere representing how close each world on the sphere is to the world that lies at the center of the given system of spheres. Thus, all worlds on a given sphere are equally similar to the world at the center, and if one sphere is inside another, then the worlds on the inner sphere are more similar to the world at the center than are the worlds on the outer sphere.

Accordingly, if one replaces the reference to a system of spheres by a direct reference to similarity, the account of the truth conditions of counterfactuals that Lewis is advancing is roughly as follows:

> The counterfactual "If p were the case, then q would be the case" is true in world W
>
> if and only if
>
> Either (1) there is no world at all where p is true, or else (2) some world W* in which both p and q are true is closer to W than any world in which p is true and q is false.

[9] David Lewis, *Counterfactuals*, p. 16.

5.2.3 A Crucial Objection to the Theory: Jonathan Bennett and Kit Fine

In his review of Lewis's book, *Counterfactuals*, Jonathan Bennett advanced a number of objections to a similarity-over-possible-worlds approach to counterfactuals. But Bennett suggested that the "fatal defect" in the whole approach was that it either generated the wrong truth-values for certain counterfactuals, or else it involved unsound judgments of similarity.

Bennett illustrated his point by a counterfactual concerning Oswald and the death of Kennedy:

> "If Oswald had not killed him, Kennedy would not have been killed."

Suppose, Bennett says, that the Warrenite hypothesis that Oswald acted alone, etc., is true. Then Lewis's approach will generate the wrong truth-value for the above counterfactual, since it is, Bennett claims, "incredible" that "some worlds in which no one kills Kennedy are more like the actual world than is *any* world in which Kennedy is killed by someone other than Oswald."[10]

Similarly, Kit Fine, in his review of Lewis's book, also argued that Lewis's approach generates the wrong truth-values for counterfactuals where the consequent could only be true if the world were radically different from the actual world, and Fine illustrated this point by the following counterfactual concerning Nixon and the button:

> "If Nixon had pressed the button, there would have been a nuclear holocaust."

5.2.4 Lewis's Response to the Crucial Objection

Lewis's response to this objection was set out in his article, "Counterfactual Dependence and Time's Arrow." The thrust of it is that the theory of counterfactuals in question need not be formulated in terms of our ordinary standards of overall similarity. Nor should it be, for then the central objection advanced by Bennett and Fine would be correct. The crucial question, accordingly, is simply whether one can specify weightings for factors that are relevant to the similarity of one world to another that will combine to produce a measure of overall similarity that will generate the right truth-values, and Lewis contends that this is possible.

In particular, Lewis proposes a weighting of factors according to which, while a perfect match of particular facts for an extended stretch of time counts for more with respect to overall similarity than the absence of a single, small, localized, miracle, or violation of a law of nature, it is less important than the

[10] Jonathan Bennett, "Counterfactuals and Possible Worlds," p. 395.

absence of large miracles. Given this weighting, any world in which Nixon presses the button, if it contains only a small miracle that stops the transmission of the signal, and thus the nuclear holocaust, will not involve a perfect match with respect to future facts–since, for example, the button will be warmer because of contact with Nixon's finger, rays of light in the vicinity of Nixon's hand will be affected differently, Nixon's memories will be different, and so on, all of which will lead to significant divergences with regard to the future state of the world at all later times. One could, of course, consider a world where all the those differences were eliminated by miracles, but then one would be achieving a perfect match at the cost of a large miracle–a cost that, according to Lewis's proposed criteria for overall similarity, is too high.

5.2.5 The "Nixon and the Button" Objection Revisited

Lewis's response enabled him to escape the specific formulations of the objection in question that were advanced by Bennett and Fine. But, as I shall now show, it cannot handle the more general, underlying objection.

The reason it cannot is that Lewis's proposed solution depends upon the fact that Nixon's pressing the button is an event which has *multiple* effects, and multiple effects of such a sort that it would take a very big miracle to remove all traces of the event, in order to make it the case that there was a perfect match with the future of the actual world. But this means that if it is logically possible to construct a parallel case where the crucial event does not have multiple effects of such a sort that it would take anything beyond a small miracle to remove all traces, Lewis's response will not work, and the fundamental objection will stand.

Can that be done? The answer is that it can be, and in at least two different ways. The first involves considering a world that contains at least one type of causal process that has the following two properties:

(1) The causal process is non-branching, at least over some temporal interval;

(2) The causal process brings about events that are causally necessary conditions for possible subsequent, branching, causal processes.

Schematically, then, the idea is that one considers a world where there are deterministic causal laws that entail that an event of type *C* will give rise to a causal process that leads, after a temporal interval t_1, to an event of type *E*, and where, moreover, at every instant during the relevant temporal interval, there is at most one event that is causally related to the event of type *C*. If an event of type *E* occurs, however, in the presence of an event of type *D*, the result will be multiply branching causal processes, leading, after a further temporal interval t_2, to the occurrence of events of types F_1, F_2, F_3, . . . F_n. Now let *W* be a world where such causal laws obtain, but where no event of type *C* occurs at time *t*, and consider the following counterfactual:

> (*) If an event of type *C* had occurred at time *t*, then events of types F_1, F_2, F_3, . . . F_n would have occurred at time ($t + t_1 + t_2$).

This counterfactual is clearly true in the world we are considering, but it comes out false on Lewis's approach. For consider two worlds, W_1 and W_2, that are otherwise as similar to the original world, *W*, as possible, but where an event of type *C* does occur at time *t*, and where the following propositions are true in the respective worlds:

In W_1: An event of type *E* occurs at time ($t + t_1$), and events of types F_1, F_2, F_3, . . . F_n occur at time ($t + t_1 + t_2$);

In W_2: No event of type *E* occurs at time ($t + t_1$), and no events of types F_1, F_2, F_3, . . . F_n occur at time ($t + t_1 + t_2$).

W_2 differs from W_1 in two respects. First, it involves a single, small violation of a law of the original world, *W*, since one has an occurrence of an event of type *C* at time *t*, but no event of type *E* at time ($t + t_1$). In this respect, W_2 is less like the original world than W_1 is. But, secondly, W_2 is a perfect match with *W* from time ($t + t_1$) onward, whereas W_1 diverges from *W* with the occurrence at time ($t + t_1 + t_2$) of events of types F_1, F_2, F_3, . . . F_n , and this divergence then becomes ever greater as the resulting causal processes continue to branch. The upshot is that, given the measure of similarity proposed by Lewis, W_2 is closer to the original world, *W*, than W_1 is, and so it follows, on his account of the truth conditions of counterfactuals, that if an event of type *C* had occurred at time t, neither an event of type *E* at time ($t + t_1$) nor events of types F_1, F_2, F_3, . . . F_n at time ($t + t_1 + t_2$) would have occurred. So counterfactual (*) gets wrongly classified as false.

Formulated in terms of Nixon and the bomb, the example could be as follows. Imagine a world that is different from ours in certain respects. First, it is a world where it is possible to bring about physical events psychokinetically. Secondly, it is a world where an act of willing that something be brought about psychokinetically involves no physical change: it consists, instead, only of an appropriate mental state involving emergent qualia. Finally, such a qualia-state is almost causally impotent: its only effect is the psychokinetically caused occurrence of the event that was willed; there is not even any memory trace of the relevant act of willing in the person who performed the act.

Here I have formulated things in terms of a direct causal connection between the act of willing and the occurrence of the event willed. If such a direct connection is thought to be somehow unacceptable, one can easily arrange a mechanism: there can be a non-branching causal chain that proceeds along a straight line to the location where the event occurs, and where no part of the intervening causal chain has any other effects.

A strange world, no doubt! Yet surely one that is logically possible. But, then, imagine Nixon–or a Nixon counterpart–in a world of this type where he does not will that the button be pressed psychokinetically. What would be the case if Nixon, in such a world, *had* willed that the button be pressed psychokinetically? The correct answer, surely, is given by the following counterfactual:

> "If Nixon had willed that the button be pressed psychokinetically, then that would have happened, and there would have been a nuclear holocaust."

But on Lewis's approach, this counterfactual will be false. For an act of willing that something be brought about psychokinetically, in the world that we are considering, will have only one effect: the occurrence of the event that was willed to happen. There is no causal branching, and so only a single, small, localized miracle is required to bring it about that although Nixon has willed that the button be pressed psychokinetically, the button is not pressed, and thus, rather than there being a nuclear holocaust, there is, instead, a perfect match with the future of the original world in which Nixon does not will that the button be pressed psychokinetically. So if a single, small, localized, miracle contributes less to dissimilarity than a perfect match with respect to all future states of affairs contributes to similarity, then it follows that the above counterfactual is false, rather than true. So Lewis's response fails: it cannot handle a variation on the Nixon and the button example.

There is a second way of showing that Lewis's response does not work, since rather than appealing to the possibility of non-branching causal processes, one can appeal instead to the possibility of causal processes that come to an end.

Schematically, the idea is that one considers a world where the occurrence of an event of type *C* gives rise to causal processes that spread out in every direction, each of which gives rise, after a distance *d*, and a temporal interval t_1, to an event of type *E*. On their own, events of type *E* have no effects whatsoever. However an event of type *E* does have effects when, and only when, it occurs together with an event of type *D*, and then the result is multiply-branching causal processes, leading, after a further temporal interval t_2, to the occurrence of events of types $F_1, F_2, F_3, \ldots F_n$. Now let *W* be a world where these causal laws obtain, but where no event of type *C* occurs at location *s* at time *t*. Assume, moreover, that there is only a single spatial location, at time $(t + t_1)$, where an event of type *D* occurs at distance *d* from location *s*. Consider, then, the following counterfactual:

> (**) If an event of type *C* had occurred at time *t* at location *s*, events of types $F_1, F_2, F_3, \ldots F_n$ would have occurred at time $(t + t_1 + t_2)$.

This counterfactual is clearly true in the world we are considering, for the occurrence of an event of type *C* would initiate causal processes that spread out in every direction, and, as a consequence, an event of type *E* would occur at time ($t + t_1$) at the one spatial location at that time where an event of type *D* occurs at distance *d* from location *s*. This, in turn, would result in the occurrence of events of types $F_1, F_2, F_3, \ldots F_n$ at time ($t + t_1 + t_2$).

Counterfactual (**) comes out false, however, on Lewis's approach. For consider two worlds, W_1 and W_2, that are otherwise as similar to the original world, *W*, as possible, but where an event of type *C* does occur at time *t* at location *s*, and where the following propositions are true in the respective worlds:

In W_1: Events of type *E* occur at time ($t + t_1$) at every location at distance *d* from location *s*, including the one such location where an event of type *D* occurs, and events of types $F_1, F_2, F_3, \ldots F_n$ occur at time ($t + t_1 + t_2$);

In W_2: Events of type *E* occur at time ($t + t_1$) at every location at distance *d* from location *s*, *except* for the one such location where an event of type *D* occurs, and no events of types $F_1, F_2, F_3, \ldots F_n$ occur at time ($t + t_1 + t_2$).

W_2 differs from W_1 in two respects. First, it involves a single, small violation of a law of the original world, *W*, since one has an occurrence of an event of type *C* at time *t*, but no event of type *E* at time ($t + t_1$) at the one location at distance *d* from location *s* where an event of type *D* occurs. In this respect, W_2 is less like the original world, *W*, than W_1 is. But, secondly, W_2 is a perfect match with *W* at every moment after time ($t + t_1$) onward, whereas W_1 diverges from *W* at every moment from time ($t + t_1 + t_2$) onward, in view of the occurrence of events of types $F_1, F_2, F_3, \ldots F_n$ at time ($t + t_1 + t_2$), and this divergence then becomes ever greater as the resulting causal processes continue to branch. The upshot is that, given the measure of similarity proposed by Lewis, W_2 is closer to the original world, *W*, than W_1 is, and so it follows, on Lewis's account of the truth conditions of counterfactuals, that if an event of type *C* had occurred at time *t*, at location *s*, then there would have been no occurrence either of an event of type *E* at time ($t + t_1$) at the one location where an event of type *D* occurs at distance *d* from location *s*, or of any subsequent events of types $F_1, F_2, F_3, \ldots F_n$ at time ($t + t_1 + t_2$). So counterfactual (**) gets classified, incorrectly, as false.

The overall conclusion, accordingly, is that Lewis's attempt to answer the most crucial objection to the whole similarity-across-possible-worlds approach to counterfactuals is unsuccessful, since the possibility of worlds that contain either a single non-branching type of causal process, or else branching causal processes that terminate after a finite time, shows that some counterfactuals get assigned the wrong truth-values by a Stalnaker-Lewis account.

5.3 Other Objections to a Stalnaker-Lewis Approach

I now want to turn to some other objections to the attempt to analyze counterfactuals along Stalnaker-Lewis lines. There are, however, a number of important objections that I shall not discuss, including the following:

(1) The Relation Between ‘P & Q’ and ‘P □→ Q’

On the Stalnaker-Lewis approach to counterfactuals it appears to be true - unless the standards for similarity turn out to be such that there can be cases where world W_2 is as similar to W_1 as W_1 is to itself - that ‘P & Q ‘logically entails ‘P □→ Q’. Jonathan Bennett argues, however, that this is clearly unsatisfactory, as is shown by cases where ‘P’'s being true makes it very unlikely that ‘Q’ is true, since in such cases ‘P □→ Q’ appears to be false even though ‘P & Q’ is true.

(2) The Relation Between ‘Would' Counterfactuals and Probabilistic Counterfactuals

Bennett’s argument suggests another objection, which is as follows. Consider an indeterministic world, and one where (a) 'P' is made true by some state of affairs, S, at time t_1, (b) 'Q' is made true by some state of affairs, T, at time t_2, (c) there is no state of affairs prior to time t_2 that is a causally sufficient condition for the existence of state of affairs T, and, finally, (d) the total state of affairs that existed at the time of state of affairs S made the probability that Q would be true equal to 0.01. Then, on the one hand, given that ‘P’ and ‘Q’ are both true, it seems that the following counterfactual is true, on a Stalnaker-Lewis approach:

(i) "If P were the case, then Q would have been the case"

On the other hand, in view of (d), it would seem that the following counterfactual must also be true:

(ii) "If P were the case, then the probability that Q would be the case would be equal to 0.01."

But aren’t these two counterfactuals logically incompatible?

(3) The "Less Complex Consequent" Objection

Consider a world where the laws are such that the following counterfactual is true:

(i) A □→ (B or C)

Suppose further, first, that the world is indeterministic, and that, in particular, while the laws entail that if A is the case, then either B or C will be the case, but do not entail that, if A is the case, then B will be the case, or that, if A is the case, then C will be the case, and secondly, that B and C are logically incompatible.

Then it would seem that neither of the following two counterfactuals would be true:

(ii) $A \;\square\!\!\rightarrow B$

(iii) $A \;\square\!\!\rightarrow C$

On Stalnaker's approach, however, the truth of (i) entails that either (ii) is true or (iii) is true, since Stalnaker's approach involves the idea that there is always a <u>closest</u> A-world. This is not so on Lewis's approach. Nevertheless, the case still poses a problem for Lewis. For suppose that the case is one where, first of all, the only difference between the closest A-worlds in which B is true and the closest A-worlds in which C is true is that B is true in the former worlds, and C in the latter, and, secondly, that the truth-maker for 'B' is a much more complex state of affairs, or a temporally more extended state of affairs, than the truth-maker for C. Then the closest A-worlds in which B is true will be less similar to the original world than are the closest A-worlds in which C is true. And so on Lewis's approach it will turn out that the counterfactual

$A \;\square\!\!\rightarrow C$

is true in the world in question. Then, since C is, by hypothesis, logically incompatible with B, the following counterfactual must also be true:

$A \;\square\!\!\rightarrow \sim B$

But surely this is wrong.

This consideration can be reinforced, moreover, if one supposes that there is a law of nature that entails not only that, if A is the case, then either B or C is the case, but that the law is a probabilistic one according to which the likelihood that B is the case is very high, and the likelihood that C is the case is very low.

(4) The "Agreement with the Actual World" Objection

This objection was set out by Pavel Tichy in his paper "A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals" (1976). Tichy formulates it in terms of a person, Jones, who always wears a hat if it is raining, whereas, if it is sunny, Jones decides, in some random fashion, whether to wear a hat or not. Suppose, then, that the world we are considering is one where it is raining on a given day, and thus one where Jones wears a hat. What is the truth-value of the following counterfactual:

> "If it had been sunny on the day in question, Jones would have worn a hat."

Since a sunny-day world where Jones wears a hat will be more similar to the rainy-day world where Jones wears a hat than will a sunny-day world where Jones does not wear a hat - other things being equal - it would seem that the above counterfactual will turn out to be true on a Stalnaker-Lewis approach. But surely this consequence is unacceptable. For given that Jones decides via a random process whether to wear a hat if it is sunny, the following 'might' conditional is surely true:

> "If it had been sunny on the day in question, Jones <u>might</u> not have worn a hat."

And similarly, if Jones decides by some process that has a 50% chance of turning out in either of two ways, then the following probabilistic counterfactual will be true:

> "If it had been sunny on the day in question, the probability that Jones would not have worn a hat would have been 0.5."

So again, the idea is that one can appeal to one's intuitions about the 'might' counterfactual and about probabilistic counterfactuals to support the idea that the following counterfactual is not true:

> "If it had been sunny on the day in question, Jones would have worn a hat."

Alternatively, one may simply appeal directly to the intuition that this latter counterfactual is not true.

The above objections seem to me plausible, and where Lewis has responded to an objection, I believe that one can show either that Lewis's response is implausible, or else that the objection can be reformulated so that Lewis's response no longer works. In the present context, however, the above four objections turn out not really to be relevant, since, as we shall see later, not just any sound objection to a Stalnaker-Lewis-style analysis of counterfactuals will do: it is crucial that the objections bear are connected with causation in a certain way. So let us turn to objections that do have such a connection.

5.2.4 The "Simple Worlds" Objection

5.2.4.1 The Case of the Single Particle World

The objection to a Stalnaker-Lewis approach to counterfactuals is based upon a type of objection which, I have argued elsewhere, applies to any reductionist account of causation, and it involves considering a world that involves only a single particle – call it 'M' – with no associated fields, gravitational or otherwise . The question then is what one is to say about the truth-values of the following two counterfactuals:

> (a) If solitary particle M had not existed at time t, then it would not have existed at any later time;
>
> (b) If solitary particle M had not existed at time t, then it would not have existed at any earlier time.

The answer is that there are two conclusions that one can draw. The first is this:

Conclusion 1: A Stalnaker-Lewis approach entails that the preceding counterfactuals must have the same truth-value.

Why so? Simply because, regardless of what factors one takes as relevant to the type of similarity that is crucial for counterfactuals, one will not be able to assign different truth values to (a) and (b) unless one assigns different weight to the temporal location of one of those factors.

But what prevents one from doing that? Mightn't one, for example, assign more weight to perfect matches in the past than to perfect matches in the future?

My answer is, first, that if one did this, then it could, I believe, be shown that the reference to overall simplicity would no longer be doing any real work, and that what one would have is what Lewis refers to in his article "Counterfactual Dependence and Time's Arrow" as 'Analysis 1' - an analysis that is framed in temporal terms. Secondly, an analysis that assigned a different weight to a factor when it was earlier than the relevant time than when it was later than the relevant time would also, I think, beg the question against time travel and backward causation. For even if one holds – as I do – that backward causation and time travel are logically impossible, this is surely not a conclusion that should be built into one's analysis of counterfactuals.

The other conclusion that one can draw is this:

Conclusion 2: Lewis's approach entails not only that the preceding counterfactuals must have the same truth-value; it also entails that they are both false, and that the true counterfactual in this situation is instead:

> (c) If solitary particle M had not existed at time t, then it (or a particle indistinguishable from it) would still have existed at all later times, as well as at all earlier times.

This is so because Lewis holds that a complete match between worlds with respect to all future events counts more for similarity than a single miracle counts against similarity, and in the single-particle world that we are considering here, it takes only a single, localized, simple miracle to bring it about that a particle just like the one that dropped out of existence at time t exists at all later times.

5.2.4.2 Lewis's Response to the "Simple Worlds" Objection

Lewis explicitly considers this sort of objection in his article "Counterfactual Dependence and Time's Arrow". Here is what he says:

"It might be otherwise if w_0 were a different sort of world. I do not mean to suggest that the asymmetry of divergence and convergence miracles holds necessarily or universally. For instance, consider a simple world inhabited by just one atom. Consider the worlds that differ from it in a certain way at a certain time. You will doubtless conclude that convergence to this world takes no more of a varied and widespread miracle than divergence from it. This means, if I am right, that no asymmetry of counterfactual dependence prevails at this world." [11]

A bold response. One is reminded once again of the saying that one person's modus ponens is another person's modus tollens. But is this 'outsmarting' maneuver at all plausible? In the first place, if one holds that there is no asymmetry of counterfactual dependence in the single particle world, then neither will there be any causation, since even if one rejects Lewis's idea that causation can be analyzed counterfactually, it is surely true that the presence of causation entails the presence of an asymmetry of counterfactual dependence. But is there any causation in the single particle universe? If one focuses upon causal interaction, it will be tempting to conclude that there isn't. But here one needs to ask, first, what account one gives of conservation laws: Are they causal laws or not? If so, then by assuming that Conservation of Mass is a law in the world we are considering, it will follow that later temporal slices of that universe are causally dependent upon earlier ones. Secondly, one also needs to ask what account should be given of identity over time. Isn't a causal analysis of identity over time very plausible? If so, how can one jettison it in the present case?

Secondly, Lewis accepts a causal analysis of temporal priority, and of the direction of time. So if the single particle world has no asymmetry of counterfactual dependence, and thus no causation, then neither can there be any states of affairs that stand in the earlier than relation. But can the single particle then be something that exists at different times? Or can it be something that has, on a persistence view, temporal parts? What can it mean to say that such a world is a temporal world, rather than a world all of whose dimensions are purely spatial ?

Thirdly, a single particle world can be viewed as having been arrived at by, so to speak, gradually removing things from a complex world. Thus, consider our own world, and consider a specific electron. Let t_1 and t_2 be any two times, where t_2 is later than t_1. Then the t_2-stage of the electron in question is caused by the t_1-stage, and, similarly, the t_2-stage of that electron is counterfactually dependent upon the t_1-stage. Now imagine how the world would be if some particle, other than the electron in question, had not existed. It would still be the case that t_2-stage of the electron in question was caused by the t_1-stage, and, similarly, that the t_2-stage of that electron was counterfactually dependent upon the t_1-stage. So consider the continuation of this process, imagining possible worlds that contain fewer and fewer particles, but still contain the electron in question. Throughout this enormously long process, it

[11] David Lewis, "*Postscripts to* 'Counterfactual Dependence and Time's Arrow'," p. 49.

remains true that t_2-stage of the electron in question was caused by the t_1-stage, and, similarly, that the t_2-stage of that electron is counterfactually dependent upon the t_1-stage. But on Lewis's view there is a point - when the next to last particle is eliminated, and one is left only with the solitary electron - when things are suddenly different: no longer is the t_2-stage of the electron in question caused by the t_1-stage, and, similarly, no longer is the t_2-stage of the electron counterfactually dependent upon the t_1-stage. Indeed, according to Lewis's view, time itself disappears, as it is no longer true that the t_2-stage of the electron is later than the t_1-stage, since the disappearance of causation entails, on a causal theory of time - which Lewis accepts -the disappearance of the earlier than relation.

5.2.5 A Causally Isolated Simple Part in a Very Complex World

Lewis's very quick dismissal of the single particle world as one that simply does not involve any asymmetry of counterfactual dependence seems, accordingly, very implausible. But what I now want to argue is that one can modify the case of the single particle world to get another case that constitutes a strong objection to the Stalnaker-Lewis approach to counterfactuals.

The basic idea is to embed the single particle scenario into a very complex world, as follows. Consider possible worlds that are rather like ours except for the fact that photons are not affected by gravitational fields. In some of these worlds there could be a single photon either that was very, very far from everything else, and that never causally interacted with anything else. Now consider the following two counterfactuals:

(a) If causally isolated photon M had not existed at time t, then it would not have existed at any later time;

(b) If causally isolated photon M had not existed at time t, then it would not have existed at any earlier time.

We can now draw conclusions precisely parallel to those that were drawn in the case of the very simple universe that contained a single particle. Thus, in the first place, we have the following conclusion:

Conclusion 1: The Stalnaker-Lewis approach entails that the preceding counterfactuals must have the same truth-value.

This obtains because, once again, the Stalnaker-Lewis approach cannot plausibly assign different weights to perfect matches which are the same, but which occur at different times.

Secondly, we have the following conclusion concerning Lewis's approach:

Conclusion 2: Lewis's approach entails not only that the preceding counterfactuals must have the same truth-value; it also entails that they are both false, and that the true counterfactual in this situation is instead:

> (c) If causally isolated photon M had not existed at time t, then it (or a particle indistinguishable from it) would still have existed at all later times, as well as at all earlier times.

And, once again, the reason is that according to Lewis's approach, a complete match between worlds with respect to all future events counts more for similarity than a single miracle counts against similarity, and in the complex world containing a causally isolated particle that we are considering here, it takes only a single, localized, simple miracle to bring it about that a particle just like the isolated one that dropped out of existence at time t exists at all later times.

A variation on this counterexample in which the particle in question is not always isolated is also possible. Thus, consider a world where photons are unaffected by gravitational fields, and where there is a photon that interacted with other particles before time t, but that did not do so at time t or afterwards, perhaps because of an unending expansion of the universe. Then, once again, the following counterfactual -

> If causally isolated photon M had not existed at time t, then it would not have existed at any later time

- will turn out to be false if one accepts Lewis's measure of similarity, since a single, small miracle, involving the existence of photon qualitatively indistinguishable from photon M at appropriate locations after time will contribute less to dissimilarity of worlds than the perfect match over all future times that is thereby achieved will contribute to similarity.

5.2.6 The Inverted Worlds Objection

This objection is based upon another objection that I have directed against reductionist approaches to causation. There the thrust of the objection is that reductionist approaches generate the wrong direction for causal processes in certain rather unusual, but logically possible universes. Similarly, the thrust here is that the Stalnaker-Lewis approach to counterfactuals generates the wrong direction for counterfactual dependence in the universes in question.

The basic idea is that there could be worlds that were 'temporally inverted twins'. To put things concretely, suppose that it is the year 4004 B.C. A Laplacean-style deity – albeit one with more of a sense of humor than many deities – is about to create a world rather similar to ours, but one where Newtonian physics is true. Having created a 4004 B.C. world, and having selected the year 2000 A.D. as a good time for Armageddon, the deity works out what the 4004 B.C. world will be like at that point, down to the last detail. He then decides to create two spatially unrelated worlds: the one just mentioned, together with another whose initial state is a 'temporally flipped over' version of the final state of the first world at the time of Armageddon. That is to say, the final state of the first world agrees exactly with the initial state of the second world, except that the velocities of the particles in the one state are exactly the opposite of the corresponding ones in the other.

Consider, now, any two complete temporal slices of the first world, A and B, where A is earlier than B. Since the worlds are Newtonian ones, and since the laws of Newtonian physics are invariant with respect to time reversal, the world that starts off from the reversed, 2000 A.D. type state will go through corresponding states, B* and A*, where these are flipped over versions of B and A respectively, and where B* is earlier than A*. So while the one world goes from a 4004 B.C., Garden of Eden state to a 2000 A.D., Armageddon state, the other world will move from a reversed, Armageddon type state to a reversed, Garden of Eden type state.

Let us refer to the two worlds as, respectively, W and W*. In W, the following counterfactual is true:

> "If state of affairs A had not obtained, then state of affairs B would not have obtained"

So if, for example, one considers Stalnaker's account, then the closest non-A-world to W - call it V - will be a non-B-world. But then, in view of the fact that W* is a temporally inverted twin of world W, the temporally inverted twin of world V - call it V* - must have the following properties:

(a) V* is a world where state of affairs A* does not obtain;

(b) V* is a world where state of affairs B* does not obtain

(c) V* is the closest non-A*-world to world W*.

So it must be case, given a Stalnaker account, that the following counterfactual is true in world W*:

> "If state of affairs A* had not obtained, then state of affairs B* would not have obtained"

And precisely the same is true if one adopts, instead, Lewis's slightly more complicated account. But the counterfactual in question is false: B* is not counterfactually dependent on A*. Rather, A* is caused by, and so is counterfactually dependent upon, B*.

In short, the Stalnaker-Lewis approach entails a consequence that is unacceptable, namely, that the following counterfactuals – the one in world W, and the other in world W* – must have the same truth-value:

> (a) If A had not existed at time t, then B would not have existed at (t + Δt);
>
> (b) If A* had not existed at time t, then B* would not have existed at (t - Δt).

5.3 A Possible Response: Causation and Stalnaker-Lewis Conditionals

The above objections provide, I suggest, decisive reasons for concluding that a Stalnaker-Lewis-style approach to counterfactuals must be abandoned. It might then seem to follow that, unless some other approach to counterfactuals that does not involve any causal concepts can be found, approaches to causation of the sort advanced by David Lewis are also doomed. But, in fact, this is not quite right.

The reason is this. Suppose that a Stalnaker-Lewis type of account of the truth conditions of counterfactuals is unsound. It is still the case that it serves to define a certain conditional - though perhaps one that does not correspond to any conditional found in any natural language. But there is surely nothing wrong with that, and so it might be suggested that causation can be analyzed, along something like the lines proposed by Lewis, in terms of what can now be viewed as a *new* conditional – the Stalnaker-Lewis conditional. The resulting analysis of causation will no longer be a *counterfactual* analysis, of course, but it need not be any the worse for that.

This is an important objection. But notice, in the first place, that it is not quite true that nothing is lost if one shifts to the view that the conditionals used in the analysis, rather than being counterfactuals, are some new type of conditional. For, after all, there do appear to be some conceptual connections between causation and counterfactuals, and even if many philosophers are inclined to think that the source of these connections lies in the fact that causation enters into the analysis of counterfactuals, rather than vice versa, the existence of such connections lends at least some intuitive attraction to the idea of a counterfactual analysis of causation. By contrast, if Stalnaker-Lewis conditionals, rather than being counterfactuals, are simply a novel type of conditional, then any intuitive basis for a Lewis-style analysis of causation would vanish, and it would then be a remarkable accident indeed if the resulting analysis turned out to be sound.

In the second place, however – and this is the crucial point – the objections on which I have focused above also serve to show that causation cannot be analyzed in a Lewis-style way. For consider, first of all, the case of Nixon and the button, and rather than a world where Nixon does not push the button, consider a world – call it W_0 – where Nixon does psychokinetically push the button, thereby producing an exciting nuclear war. In that world, what would have happened if Nixon had not willed that the button move? One possibility is this:

W_1: The button does not move, and there is no nuclear war, but the world is otherwise as close to W_0 as possible.

Another possibility is this:

W_2: The button does move, there is a nuclear war, and all events from the time at which the button moves, on into the future, agree completely with those in W_0.

Which of these worlds is closest to W_0? W_2 matches W_0 perfectly with regard to all future events after a certain time, whereas W_1 does not. On the contrary, the future of W_1 is radically different from that of W_0. On the other hand, W_1 does not involve any miracles at any time after the moment at which Nixon decides not to push the button psychokinetically, whereas W_2 does, since the button moves, even though it was not psychokinetically pushed. But the miracle in question is only a single, small miracle, and, according to Lewis's criteria for similarity of the relevant sort, the presence of a complete match of W_2's future with that of W_0 is a more significant factor than the fact that W_2 involves a single, small miracle, while W_1 does not. Therefore, world W_2 is closer to world W_0 than world W_1 is. Consequently, the following Stalnaker-Lewis conditional will be true in world W_0:

> If Nixon had not willed that the button be pressed psychokinetically, the button would have moved.

Thus the movement of the button is not Stalnaker-Lewis, conditionally dependent upon Nixon's willing that the button be pressed psychokinetically, and since such dependence lies at the very heart of the approach to causation that we are considering, it follows from the truth of the above conditional that Nixon's willing that the button be pressed psychokinetically does not, in world W_0, cause the button to move. But, by hypothesis, it does. Therefore causation cannot be analyzed in terms of the novel, Stalnaker-Lewis conditionals.

The other objections on which I focused lead to the same conclusion. Consider, for example, the case of the isolated particle in the very complex world. Whatever measure of similarity one chooses, the two conditionals we considered earlier will have the same truth-values, and so it will be the case either that later temporal stages of the particle are not causally dependent upon earlier ones, or that they are, but that it is also the case that earlier temporal stages are also causally dependent upon later ones, and neither consequence seems acceptable. In addition, if one adopts the measure of similarity that Lewis advances, it will turn out that if the particle had not existed at any moment, it would still have existed at later moments, and so later stages of the particle turn out not to be causally dependent upon earlier stages.

Finally, consider the inverted world case. Whatever measure of similarity between worlds one chooses, if that measure generates Stalnaker-Lewis, conditional dependence that runs in the right direction in the non-inverted world, it will generate conditional dependence that runs in the opposite direction to that of causation in the inverted world. So once again, any analysis of causation in terms of Stalnaker-Lewis conditionals cannot be sound.

5.4 Summing Up: Counterfactual Analyses of Causation

A number of philosophers have advanced some very important objections to counterfactual analyses of causation – objections that I surveyed very briefly at the beginning of this paper. These include: (1) the dependence of Lewis's analysis upon an account of the individuation of events; (2) the existence of counterfactuals that have nothing to do with causation, or even with laws of nature; (3) counterfactuals that are based upon non-causal laws; (4) situations involving causal overdetermination; and (5) cases of preemption, including 'trumping' preemption.

All of these objections pose serious, prima-facie obstacles for any counterfactual analysis of causation. It may well be, however, that some of them can be surmounted by an appropriately formulated account. Thus I am inclined to think, for example, that difficulties concerning the right account of the individuation of events can be eliminated simply by adopting the view that causal relata are states of affairs, rather than events. On the other hand, some of the other objections seem much more problematic, and, in particular, it seems to me very doubtful that there is any satisfactory way of handling the causal overdetermination objection.

The basic thrust of this discussion, however, has been that there is a more fundamental flaw in the whole idea of a counterfactual analysis of causation, and one which cannot be avoid by any tinkering with the details of the analysis. This is, first, that a counterfactual analysis requires an account of the truth conditions of counterfactuals that does not itself involve any causal concepts, and secondly, that there are excellent reasons for thinking that no adequate, non-causal account of counterfactuals is at all likely to be forthcoming, since the only serious candidate for such an account – namely, a Stalnaker-Lewis-style analysis of counterfactuals – appears to be open to a number of decisive objections.

Finally, I considered the idea that one might abandon the project of a *counterfactual* analysis of causation, but hold that causation can still be analyzed in terms of Stalnaker-Lewis-style conditionals, where the latter are no longer identified with counterfactual conditionals. Such a move would, I noted, deprive the proposed analysis of any intuitive basis. But, more importantly, one can show that a number of central objections to a Stalnaker-Lewis approach to counterfactuals also tell against any attempt to analyze causation in terms of new, non-counterfactual conditionals defined along Stalnaker-Lewis lines.