tions? Why do we regard *earlier* conditions as more basic than later ones. At least part of the answer lies, I think, in the fact that earlier sets of circumstances are simpler (in a certain respect) than later ones. This is due to the familiar fork asymmetry—correlated, simultaneous events are linked to a characteristic prior event that determines them both, but there need be no characteristic subsequent event. Consequently correlations can be derived from a more unified earlier condition. So there is a gain in the simplicity of our characterization of the world as we move back in time from states of correlation to their unified determinants. In other words, the earlier, central event in a fork allows a unified derivation of the separated correlated events and is therefore explanatorily more basic than them. However, this is not the whole story. A complete account of the direction of explanation would have to mention a *cluster* of factors, paralleling those maxims invoked earlier to explain our faith in the direction of causation. Specifically, we might say that we simply believe that *earlier* events are more basic than *later* ones, that the known seems 'more real' and more basic than the unknown, and that our decisions are more basic than the events that are nomologically connected to them. It's not clear that we can deny any of these maxims a role in fixing the direction of explanation.

In chapter 8 I sketched a construction of the causal relation that began with the idea of 'chains of direct nomological determination', and then added constraints of continuity and typical time order to provide the chains with a definite causal interpretation. I argued that each elementary link in these chains is an instance of an explanatorily basic law of nature. And now we have just seen that the time-order constraint may also be derived from a view of what is explanatorily basic. These considerations suggest that we can characterize the nature of causation from an independently definable notion of explanation, but not vice versa. Thus there is reason to suppose that 'explanation' is theoretically prior to 'causation'.

# 10
# Counterfactuals

## 1. *Goodman's problem*

Counterfactuals concern what would happen, or have happened, in specified hypothetical circumstances. For example, one might say of an unused match, "if struck, it would have lit". The analysis of such statements is important. For, in the first place, they are pervasive: the meanings of countless scientific predicates (soluable, malleable, ductile, explosive, etc.) appear to involve commitments to what would occur under certain possible circumstances. And second, the notion of counterfactual dependence is bound up with other puzzling concepts, such as 'law', 'cause', and 'explanation'. Yet, despite considerable attention by philosophers of science in the last forty years, there has emerged no satisfactory way of describing what must be the case for such conditional statements to hold. In particular, it is evident that the truth of "If $p$ were true, then $q$ would be true" is not determined solely by the truth values of $p$ and of $q$. Typically $p$ and $q$ are both believed to be false, yet it remains open whether or not the counterfactual claim is acceptable.

One of the first sustained attempts to grapple with this matter is Nelson Goodman's classic essay, "The Problem of Counterfactual Conditionals" (1946). He suggests, there, that a counterfactual, $p \mathrel{\Box\!\!\rightarrow} q$ is true when its antecedent, $p$, nomologically requires, given prevailing conditions, the truth of its consequent, $q$. Or, in other words, that the antecedent, $p$, conjoined with some set of facts, $S$, and laws of nature, $L$, deductively entails the consequent, $q$:

$$p \mathrel{\Box\!\!\rightarrow} q \quad \text{if and only if} \quad (p \,\&\, S \,\&\, L) \rightarrow q$$

However, this idea has foundered on two major obstacles: the so-called "cotenability problem", and the explication of "laws of nature". I shall consider the first of these difficulties immediately, and postpone the second one until section 4.

It is evident that some restriction must be placed on the contents of $S$. For suppose the antecedent is in fact false, as in the case of the

unstruck match. And suppose that not-$p$—which, in that case, is a fact—were to be included among the prevailing conditions, $S$. Then $(p \,\&\, S \,\&\, L)$ would be a contradiction, entailing any $q$ whatsoever. In this way all counterfactuals (with false antecedents) would be made trivially true.

In the attempt to overcome this sort of difficulty, it turns out that membership in $S$ has to be restricted to facts that would not be altered by the truth of $p$. That is to say, $S$ must be required to satisfy the condition: $p \,\Box\!\!\rightarrow S$. Thus Goodman is forced to acknowledge that the restriction clause must itself involve counterfactual conditionals, thereby rendering the whole analysis circular. For example, we accept

If the match had been struck, it would have lit

because the match was dry and oxygen was present. We do not, however, accept

If the match had been struck, there would have been no oxygen

even though the match was dry and did not in fact light. Goodman found, as I have just said, that he could match these intuitive judgments only by requiring that the background facts, $S$, be *cotenable* with the antecedent (i.e., be such that they would remain true even if the antecedent were true). Given such a rule, the non-lighting of the match cannot be invoked as one of the background facts. But, as Goodman points out, this solution is no good at all, since it relies upon the very notion we were trying to explain.

## 2. *A causal/explanatory theory of counterfactuals*

Can we use the notion of causation to solve this problem? Causation and counterfactual dependence are intimately related ideas, and it is natural to want to explain one in terms of the other. This has been tried by Lewis (1973b), who analyzes causation in terms of counterfactual dependence. However, I believe that his approach puts things the wrong way round and that this basic error yields a host of counterintuitive consequences. Happily, there is a decent alternative. Our predetermination-chain theory of causation, based on the concept of law, is at hand (see chapter 8). If this account is roughly right, then there is no need for a prior grasp of counterfactuals. On the contrary, the way is clear to use causal concepts in specifying the meaning of counterfactual conditionals. Let me now indicate how this can be done. In the final section of this chapter, I shall give extra support to this approach by describing and criticizing Lewis's theory in some detail.

My plan is to use causation in a solution to the cotenability problem, thereby enabling a revival of something like Goodman's strategy. He proposed that the counterfactual $p \,\Box\!\!\rightarrow q$ is true just in case the truth of $p$ would, *given the circumstances* $S$, nomologically determine the truth of $q$. But, as we have seen, this theory runs into trouble over the word "circumstances". For it looks as though the only adequate account of that notion would itself employ counterfactuals, making the analysis viciously circular.



I want to suggest, however, that the relevant conception of 'circumstances' may be described in terms of causation, rather than counterfactual dependence. Specifically, we should consider the circumstances $S$, (in which the truth of the antecedent is to be supposed) to consist of any facts that are *not* causes of, caused by, or nomologically determined by $-p$, the falsity of the antecedent. Thus, in order to validate $p \,\Box\!\!\rightarrow q$, imagine a change in the world ($p$ instead of $-p$), hold fixed things that are logically, nomologically, and causally independent of the changed fact ($-p$), and observe that the consequent ($q$) would be determined. In other words, consider a possible world containing all phenomena that are neither causes nor effects of (nor nomologically determined by) the events described by $-p$ but containing $p$. If (and only if) our laws of nature determine that $q$ obtains in that world, then the counterfactual, "If $p$ were the case, then $q$ would be the case", is definitely true. If our laws entail merely a certain probability, $x$, that $q$ obtains, then the probability is $x$ that the counterfactual is true. I think that this resolves the cotenability problem—and without a hint of circularity since counterfactuals are not themselves relied on in our account of causation.

Consider, for example, the case of the unused match. How can we accommodate the fact that we accept

If the match had been struck, it would have lit

but not

If the match had been struck, there would have been no oxygen

The answer, in terms of our causal theory, is obvious. The actual nonstriking of the match caused neither the presence of oxygen nor the dryness of the match. And it was not caused *by* either of them. Therefore both facts qualify to be among the circumstances of the supposed striking, and so the lighting will indeed be determined. On the other hand, the nonstriking of the match *was* a cause of its not lighting. Therefore the nonlighting fails our condition for inclusion among the circumstances of the supposed striking. Thus, although the nonlighting would, given the antecedent, imply an absence of oxygen, this fact does not validate the second conditional.

In the present account, no hypothetical *particular* fact can counterfactually imply the violation of a law of nature. This is because in determining what depends counterfactually on an event $E$, the theory tells us to hold fixed anything that is logically, nomologically, or causally independent of $E$, and any law of nature will meet that condition. Thus there are no true counterfactuals of the form. "If $E$ had not occurred, then $L$ would be false", where $E$ is a particular event and $L$ is a law of nature. However, the theory does not exclude *counterlegals*, which are counterfactuals whose antecedents hypothesize the falsity of laws. In such a case we must hold fixed any event whose occurrence was not governed by the laws whose violation is supposed, as well as any further laws that neither explain nor are explained by them.

Notice that my characterization of the meaning of counterfactuals is given in terms of 'assertibility conditions' rather than truth conditions. In other words, instead of providing a traditional style of analysis for "$p \mathrel{\square\!\!\rightarrow} q$", I have specified the circumstances in which one ought to believe to degree $x$ that $p \mathrel{\square\!\!\rightarrow} q$. That is to say, I have described the conditions for

$$\text{Prob}\,(p \mathrel{\square\!\!\rightarrow} q) = x.$$

My reason for taking this approach is an inclination to agree with Robert Stalnaker (1984) that $p \mathrel{\square\!\!\rightarrow} q$ might be true even if $p$, given the prevailing circumstances, would certainly not determine $q$. I would say, for example, that there is a probability of one half that if a certain fair coin had been tossed it would have landed heads up. But in these cases (when they are genuinely random) there appear to be no facts, more basic than the counterfactuals themselves, in virtue of which they are true. Let me emphasize, however, that even if I am wrong on this matter, the main point remains unaffected: namely, that the problem of specifying 'prevailing conditions' need not introduce a circularity into the analysis of counterfactuals, since it may be solved by employing the concept of causation.

Notice also that my proposal is specifically tailored to counterfactuals that concern events. In order to obtain a more general account, we must generalize the cotenability condition. Instead of supposing that the 'prevailing conditions' be facts that are not *causally* explained by, or *causal* explanations of, the falsity of the antecedent; we must drop the restriction to *causal* relations and say, more generally, that the 'background facts' are facts that are not explanations of, or explained by $-p$—in whatever sense of explanation is appropriate to the domain in question. Thus, it would be more accurate to call our account "an explanation theory of counterfactuals". For causation enters the picture only because, in the domain we are mainly interested in, explanation is a specification of causes.



### 3. *Is counterfactual dependence time-asymmetric?*

One big difference between the preceding account and conventional wisdom about counterfactuals is over the alleged *direction* of counterfactual dependence. It is often said that true counterfactuals are typically future-oriented, concerned with what would happen *subsequently* if some hypothetical event were to occur. Indeed, this alleged feature of counterfactual dependence is essential to theories, such as Lewis's, that analyze causation in terms of counterfactual dependence and that attempt to explain the direction of causation in terms of the temporal properties of counterfactual dependence.

Proponents of this view will generally concede that there are true past-oriented counterfactuals like

If the match had lit, it would have been struck

But they claim that in such cases a *nonstandard* notion of counterfactual dependence is employed (deriving, in Lewis's theory, from a nonstandard metric of possible world similarity). On our account, however, the counterfactual "if. . . then. . . " is univocal, and a fact may depend on a later event just as easily, and in the same sense, as it depends on an earlier event. Thus on the present account counterfactual dependence is not asymmetrical with respect to time.

This thesis might seem wrong. For one must admit that it is unusual to find a statement of the form

If $p$ were true, then $q$ would have been true

when $q$ is about events earlier than the time of $p$. However, the difficulty here disappears as soon as we recognize that past-oriented counterfactuals are only rarely, and with considerable strain, expressed in precisely that form. Instead, we tend to say

If $p$ were true, then $q$ would *have to* have been true

or more often,

$p$ would be true, only if $q$ had been true

Thus future- and past-oriented counterfactuals tend to be formulated in different ways. Nevertheless, the underlying relation of counterfactual dependence is time-asymmetric.

In denying that there is any need to proliferate senses of the counterfactual conditional, I follow the lead of Jonathan Bennett (1984).

He points out that the temptation to postulate an ambiguity comes from the existence of pairs of seemingly acceptable conditionals whose apparent incompatibility may be dissolved by acknowledging an equivocation. For example, there are circumstances in which both

1. If he had jumped off the building, he would have been hurt

and

2. He would have jumped, only if there had been a safety net

would seem to be true. And a way to reconcile them is to suppose that different kinds of dependence are involved.

However, as Bennett rightly says, the intuition that both statements are true is not strong enough to justify a substantial increase in theoretical complexity. Rather than introduce multiple senses of the conditional, we can simply insist that in fact, depending on the circumstances, only one of those claims is true. For example, if the man were on the verge of jumping and was barely talked out of it, then statement 1 looks plausible. But if he merely went up on the roof to admire the view and was jokingly challenged to jump, then 2 seems like the right thing to say. Moreover we can soften the conflict with our initial intuition even further, by recognizing that the *full* antecedent of a counterfactual conditional may not always be made explicit. Thus someone might say 1 when what he means is, "If he had jumped and there had still been no safety net, then he would have been hurt"—thereby reconciling the thought behind 1 with 2.

This kind of strategy is taken to extreme lengths in a theory of counterfactuals propounded by van Fraassen (1980). According to his view, a counterfactual conditional is simply a logical entailment claim whose antecedent is only partially explicit and residually tacit. That is,

$$p \mathrel{\square\!\!\rightarrow} q \quad \text{if and only if} \quad (p \,\&\, F) \rightarrow q$$

where $F$ is a body of facts whose content is decided on by the speaker but not expressed. Thus in different contexts, depending on what the speaker has in mind, a single counterfactual may have different truth values. Moreover this context dependence is taken to imply that counterfactual statements are not scientifically objective.

In response to this idea, I would suggest that van Fraassen overestimates the extent to which a speaker is required to dictate the circumstances in which his antecedent is to be supposed. The speaker does not need to intend that the circumstances include facts that are causally independent of $-p$, for this is already implicit in the meaning of the conditional. Moreover, if our linguistic practice were as van



Fraassen says it is, then we would be hard pressed to explain disagreement over counterfactual theses. Surely these conflicts of opinion are not all simply a result of failure to appreciate what the speaker is assuming. Moreover, the disagreements are typically settled by reference to empirical evidence; and again this would be inappropriate if counterfactuals were compressed assertions of logical entailment.

This is not to say that the speaker's intentions have *no* role to play. As we have seen, they can determine which particular counterfactual statement is expressed by the utterence. But radical conclusions do not follow from this concession. First, it does not transform counterfactuals into entailment claims. And second, it renders counterfactuals no less scientifically objective than other theses that are not always spelled out explicitly.

### 4. *Laws of nature*

Besides the question of cotenability a further difficulty with any Goodmanesque view of counterfactuals, such as the one I am proposing here, is the need to explicate the notion of *law*. For a counterfactual definitely holds only when its consequent *must*, by law, obtain, given the circumstances. But what are we saying when we maintain that some statement does not just *happen* to be true, but is a law of nature? What distinguishes laws, such as

All emeralds are green

from non-laws, such as

All the coins in my pocket are dimes

We must confront this question, not only for the sake of understanding counterfactuals, but also in order to ground the analysis of causation in terms of law, that was given in chapter 8.

To begin with, Goodman points out that no purely syntactic criterion will distinguish laws from other facts. Both of the preceding statements, for example, are universal generalizations. Moreover any particular fact, "$k$ is an $F$", is logically equivalent to the syntactically general, "Everything identical to $k$ is an $F$". Instead, Goodman proposes an epistemological answer derived from Hume. What is special about laws, he says, is their role in enabling us to make predictions.

This idea reflects the analogy between projecting a generalization onto purely *hypothetical* entities, and projecting it onto *unobserved, actual* entities. Suppose that statements that serve one purpose also serve the other. Then the condition for being lawlike (i.e., a law if true), and thus capable of sustaining counterfactuals, will be a suita-

bility for making predictions. In particular, a generalization of the form "All $A$'s are $B$" will count as a law, if and only if it is projectible—that is, the observation of $A$'s that are $B$ provides reason to believe that unobserved $A$'s are $B$.

But when is this so? The problem of distinguishing projectible and nonprojectible generalizations—called by Goodman, "the new riddle of induction"—is too large and difficult a topic to discuss here. However, a plausible preliminary view of the matter (Goodman 1955; Horwich 1982) is that hypotheses are projectible when they employ terms that are *natural* (that is to say, entrenched words such as "green" rather than 'defined' words such as "grue"), and when they are nevertheless syntactically *simple* (like "All $A$'s are $B$", and unlike "All $A$'s are either sampled and $B$ or unsampled and $-B$"). Given some such account of projectibility, Goodman's account of laws reduces to the idea, roughly speaking, that laws are naturally simple generalizations.

I think we should accept the Humean idea that there is an intimate relationship between lawlikeness and being a naturally simple, projectible generalization. But it oversimplifies matters simply to identify these notions. To see this, notice that non-laws are often projected. For example, suppose that a box of matches is dropped in a puddle, and there is then some question as to whether they have been ruined. I try a few from various parts of the box, and all of them light quite easily. On this basis I confidently conclude:

Every match in the box is dry

which is clearly not a law of nature but nevertheless projected from a small sample. Similarly consider Hempel's (1966) notoriously troublesome example:

All objects made of pure gold have a mass of less than 100,000 kg

This isn't a law. But its high credibility does not depend on the observation of every single gold object.

Thus there are true, projectible generalizations that are not laws of nature. And this should not be surprising. All it takes for a generalization to be projectible is the belief that its truth would not just be a coincidence, but that there would be some uniform reason for conformity. In other words, a projectible generalization cannot be wholly accidental. But there are many generalizations that are true because of some combination of laws and particular fact. We might call them "semi-laws". In such cases the nomological component may render them projectible; yet the component of particular fact will disqualify them as pure laws of nature.



In response to this objection it might be said that the alleged counterexamples are not *absolutely* projectible generalizations (confirmable *solely* by the discovery of positive instances) but rather merely *relatively* projectible (confirmable by positive instances, only given the right background of further assumptions). For example, the discovery of dry matches confirms the hypothesis that all the matches are dry only relative to a context in which it is known that the box was dry initially, then dropped into a puddle, and so on. So "All the matches are dry" is merely relatively projectible. Therefore one might argue that lawlikeness should be identified with absolute projectibility, so the alleged counterexamples will be disqualified. However, this refinement of Goodman's idea cannot be right. For, whether or not a hypothesis $H$ is absolutely projectible is an a priori matter, whereas whether or not $H$ is lawlike is a posteriori. Let me elaborate. Most evidence claims, in ordinary language, are dependent for their truth or falsity on the theoretical context in which they are made. Thus a particular claim may reasonably be asserted at one time and then denied later on, given a background of new theoretical beliefs. However, the possibility of such variation is eliminated in the case of evidence claims that explicitly specify their theoretical background. In other words, whereas "$E$ confirms $H$" may be true when asserted at one time and false when asserted later, because of a change of theoretical background from $B1$ to $B2$, the statements "$E$ confirms $H$ relative to $B1$" and "$E$ does not confirm $H$ relative to $B2$" are not subject to such variation. In particular, if we let $B0$ represent the 'null' background, in which no further facts are presupposed, then "$E$ confirms $H$ relative to $B0$" is not subject to revision in light of new discoveries. Thus absolute projectibility is an a priori matter. But lawlikeness is not. It is perfectly reasonable to believe at one time that $H$ is lawlike (i.e., that if it is true, then it is a law) but subsequently in the light of new information to reach the conclusion that even though $H$ is true, it is *not* a law. For we might come to think that $H$'s truth would be in part a consequence of some quite accidental particular fact. Thus whether H is lawlike depends to some extent on which other theories are true. Therefore we cannot identify lawlikeness with absolute projectibility.

This discussion indicates that besides the notion of '*naturally simple (projectible) generalization*', a further ingredient of the concept of law is *explanation*. More specifically, it seems that in order to arrive at the set of laws, we must somehow restrict the class of naturally simple generalizations; and a plausible way to do so is by means of the following 'explanation requirement':

> Laws are explainable, if at all, only in terms of other laws—and never in terms of facts that arn't laws.

Assuming an understanding of explanation (from chapter 9), we can use this principle to whittle down the set of true, simple generalizations, throwing out any whose explanation involves particular facts or other nonlaws. We will be left, at the completion of this procedure, with the laws of nature.

Here are some examples that illustrate how the explanation requirement works. By this principle the fact that objects at sea level fall with an acceleration of 32 ft/sec/sec is not a law, since its truth depends on the particular fact that the Earth has a certain mass. Nor are Kepler's so-called laws really laws, since their explanations hinge not only on Newton's laws of motion but also on the fact that none of the planets is sufficiently near to one another, or sufficiently big, to significantly distort their orbits from the type of path that Kepler characterized. On the other hand, in a classical world Newtonian mechanics contains nothing but laws, since its generalizations are not explainable at all—let alone by particular facts. And consequently, Boyle's law, relating the volume and pressure of an ideal gas at constant temperature, is a law insofar as it can be explained solely by the laws of mechanics.

Of course this outline of an approach raises many questions. Nevertheless, sketchy as it is, it does enable us to clarify certain issues. A singular virtue is that it provides a rationale for excluding Hempel's example

> All objects made of pure gold have a mass of less than 100,000 kg

from the class of laws of nature. This problem has proved surprisingly difficult. First, it plainly won't help to insist that laws be general. For the generality of this example is no mere trick of syntax. Second, as we have seen, Goodman's projectibility criterion will not suffice. On the contrary, this is one of the counterexamples to his account. And third, Hempel's own proposal will not do. He suggested that the example could be excluded because, if admitted as a law, it would preclude certain phenomena that our current theories allow to be perfectly possible. But surely that rationale cannot be correct. It would dictate that we accept *no* new laws, since any new law, unless it logically followed from already accepted laws, would be bound to narrow our view of what is possible.

However, from the perspective of the 'explanation requirement', it is not hard to see why Hempel's example is not a law. That hypothesis would not be among the set of projectible, true generalizations remaining after all those whose explanation involves non-laws are thrown out. Rather, its explanation clearly would involve elements from the set of nonprojectible facts. For example, there would be a need to cite details concerning the psychological states of the controllers of huge resources, and the absence of any motive for setting about to falsify the generalization. In short, it isn't a law because it is true in virtue of particular facts.



### 5. *Lewis's program*

In reaction to the two classic difficulties in Goodman's treatment of counterfactuals—the cotenability problem and the explication of law—a radically different approach was instigated by Stalnaker (1968, 1984) and has been developed by Lewis into a broad account of temporally asymmetric phenomena. I would like to end this chapter by looking carefully at Lewis's theory. We shall find that it is faced with a variety of criticisms—avoidable, if at all, only at the cost of cumbersome, ad hoc modifications. Therefore we shall get further support for the point of view developed earlier—in which counterfactual conditionals are analyzed in terms of causation.

Lewis's theory of causation and counterfactual dependence splits into four stages. In the first place, he (1973b) says that causation consists in a chain of counterfactual dependence:

> $C$ caused $E$ *if and only if* there was a sequence of events $X1$, $X2$, . . . , $Xn$, such that:
>
> if $C$ had not occurred, then $X1$ would not have occurred, if $X1$ had not occurred, then $X2$ would not have occurred, . . .
> if $Xn$ had not occurred, then $E$ would not have occurred

(Lewis 1987 subsequently generalizes the account to accommodate indeterministic causation.)

Second, this analysis is supplemented (1973a) with a semantic theory of the counterfactual conditional:

> If $p$ were true, then $q$ would be true *if and only if*
> there is a possible world in which $p$ and $q$ are true that is more similar to the actual world than any possible world in which $p$ is true and $q$ is false

Third, Lewis (1979b) fills out the picture with an account of the features of possible worlds that make them more or less similar to actuality. The most similar worlds are said to be those in which our laws of nature are rarely violated. But exact similarity with respect to

particular facts in some large region of spacetime is also a major factor and will promote similarity even at the cost of minor violations of law. A significant element in this account is that there is no built-in time bias. Concepts of temporal order are not employed at all in the principles describing the determinants of similarity.

Finally, Lewis's (1979b) fourth assumption introduces the time asymmetry that provides the ultimate basis for the directionality of counterfactual dependence and hence of causation. He makes the empirical claim that (almost) every event is grossly overdetermined by subsequent states of the world, but is not so overdetermined by its history. Or, in other words, that the future of every event contains many independent definite traces of its occurrence, although beforehand there need have been little or no conclusive indication that it would happen.

These four ingredients work together as follows. Let us imagine a hypothetical change in the course of the world—specifically, that some actual event $C$ at time $t$ did not occur at that time. It would be hard to reconcile this supposition with what actually happened after $t$, for in fact $C$ brought about many phenomena that determine that $C$ did occur at $t$. On the other hand, it would be relatively easy to square the supposition that $C$ did not happen with the course of the world before $t$. Because although that history may have determined that $C$ would occur, events are not substantially overdetermined by what preceded them. Thus at the cost of a small violation of our laws of nature, we can reconcile the nonoccurrence of $C$ with the actual history of the world before $t$. But we cannot, without much greater cost, reconcile this with the actual future of the world after $t$. Consequently, among possible worlds without $C$, those that are just ours until $t$ and then diverge are more similar to the actual world than those that are just like the actual world after $t$, or those that differ from the actual world before $t$. Thus from the asymmetry of overdetermination it follows that if the present were different from the way it is, then the future would be different, but not the past. Counterfactuals of the form, "If $C$ had not occurred, then $E$ would not have occurred", where $C$ was later than $E$, will be false. Consequently there will normally be no chain of counterfactually dependent events leading backward in time. Therefore effects will not precede their causes.

Difficulties with Lewis's theory of causal direction emerge at each of the four stages. Let us consider some of these problems beginning with objections to the very idea that causation should be analyzed in terms of counterfactual dependence, regardless of how such dependence is itself to be construed.



1. *Causal overdetermination.* This occurs when an event is the product of more than one causal chain which would each have been sufficient to produce the event. For example, a man's death may be causally overdetermined if he is shot in the head simultaneously by two people Smith and Bloggs, acting independently of one another. In such a case the effect is not counterfactually dependent on its causes. The man would have died even if Smith had not shot him. Nevertheless, I think we would say that Smith's shot was a cause of his death. Therefore, contrary to Lewis's analysis, the presence of a chain of counterfactual dependence is not necessary for causation.

Lewis (1973b) is perfectly aware of such cases but does not regard them as counterexamples to his view. For he believes that it is unclear how to apply causal terminology to instances of overdetermination. However, even if he is right about this (which seems doubtful), it is still a mark against his analysis that it yields a *definitely* negative answer to the question of whether Smith's shot was a cause of death. For if our conception of causation neither clearly applies nor clearly fails to apply, then an accurate analysis should reflect this indeterminacy. Note that no such difficulty with overdetermination confronts our causal theory of counterfactuals. That theory will correctly *deny* that if Smith had not fired, the victim would still be alive. For, since the shots were causally independent of one another, the occurrence of Bloggs's shot will be among the circumstances in which the absence of Smith's shot is supposed.

2. *Noncausal determination.* A counterfactual dependence between events is often associated, as Lewis says, with a causal relation between them. But it need not be. There are other alternatives (Kim 1973; Sanford 1976). For example,

> If John had not been killed, his wife would not have been widowed
>
> If the last chapter had not been written, the book not have been completed

Thus, as our analysis allows, counterfactual dependence does not imply causal connection.

3. *Directionality.* Even when the counterfactual dependence of $E$ on $C$ *does* reflect some sort of causal connection between them, this need not be because $C$ causes $E$. As we have seen, it may be, rather, that $C$ is an *effect* of $E$. For example,

> If the match had lit, it would have been struck
> If I had jumped, there would have been a net outside the window

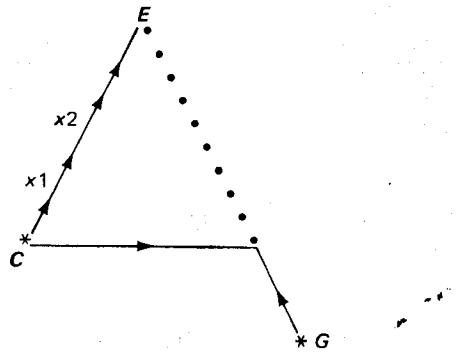


Figure 34

To handle this problem, Lewis is forced to postulate a special "backtracking" sense of counterfactual dependence, associated with special rules for measuring the similarity of possible worlds. But I argued in section 3 that there is little pretheoretical rationale for this multiplication of senses.

4. *Causal preemption.* This takes place when the cause of an event prevents something else from causing that event. For example, Smith's shooting a man preempts Bloggs's shooting him if Bloggs is frightened off before firing by the sound of Smith's gun. If $C$'s causing $E$ preempts $G$'s causing $E$, then, on the face of it, $E$ is not counterfactually dependent on $C$ because, even if $C$ hadn't occurred, $E$ would have been caused by $G$ instead. Thus preemption might seem to present a problem for the counterfactual theory of causation.

But, as Lewis points out, his analysis can nevertheless be satisfied, for there may be a chain of causal dependence ($x1$, $x2$, . . . ,) connecting $C$ and $E$, as shown in figure 34. One might still be tempted to deny that $E$ depends on $x2$, arguing that if $x2$ had not occurred, then neither would $x1$ nor $C$, so $G$ would not have been preempted from causing $E$. However, says Lewis, since we are not employing 'backtracking conditionals', $x2$'s nonoccurrence would not have made any difference to prior members of the chain; so $C$ would still have been there and would have still preempted $G$'s causing $E$.

One way of criticizing this strategy is to repeat the complaint made in point 3 regarding the alleged distinction between normal and backtracking conditionals. Another objection emerges if we consider cases where $C$ causes $E$ *directly*—without there being any intermediate event $X$ such that $C$ causes $X$ and $X$ causes $E$. If we now suppose that $C$ preempts $G$'s causing $E$, we then have a case in which Lewis's escape will not work, and his counterfactual analysis breaks down.



Suppose, for example, that ball $A$ rolls into ball $B$, causing $B$ to move out of the way of ball $D$, which would have caused exactly the motion of $B$ that $A$ actually causes. This is a case of preemption. $A$'s striking $B$ causes $B$'s motion; however, if $A$ had not struck $B$, $B$ would nevertheless have been set into motion by $D$. Therefore $B$'s motion is not counterfactually dependent on A's hitting $B$. Moreover there are no events that mediate the causal connection between $A$'s hitting $B$ and $B$'s motion. So there is no event that depends on $A$'s hitting $B$ and that $B$'s motion depends on. Thus Lewis's counterfactual condition seems to be too strong.

Again, the alternative approach is not subject to this difficulty. When $C$ preempts $G$ from causing $E$, our theory will say, as it should, that if $C$ had been absent, then $E$ would have been caused by $G$. Moreover the account of causation in chapter 8 entails correctly that $G$ is not actually a cause of $E$. For a cause must be *essential*, given surrounding circumstances, for the determination of its effect. But the only antecedent conditions that in combination with $G$ will determine $E$ are conditions that include $C$ (or its causes or effects), and such conditions determine $E$ without the help of $G$.

Admittedly, none of these arguments constitutes a knockdown argument against Lewis's approach. With enough cleverness it will no doubt be possible to save the theory from counterexamples. Indeed, Lewis (1987) does have ingenious ways of elaborating and extending his approach to deal with the problems just discussed. However, one cannot help but have the sinking feeling that we are heading for an interminable series of objections and modifications, and that even if there is an end result, it will not have the simplicity and intuitive appeal that recommended the original version. Thus, even though no irrebuttable objection to Lewis's program may be at hand, there are grounds for dissatisfaction, and reason to cast around for an alternative. These concerns are compounded as we go on to consider objections to the further stages of Lewis's theory. Let us now look at a problem that emerges when the analysis of causation is supplemented with the theory of counterfactual conditionals.

5. *Psychological implausibility.* According to Lewis, a counterfactual holds when the consequent is true in possible worlds very like our own except for the fact that the antecedent is true. But it is vital that the degree of similarity not be assessed by intuitive pretheoretical criteria. Rather, the relative importance of various factors in determining how similar some world is to our own must be retrieved from our views about which conditionals are true and which are false. For example, it has often been objected against Lewis that on his view

If the president had pressed the button, a nuclear war would have ensued

must be false, since a world in which the circuit fails and there is no war would be more like actuality than a world in which all life is destroyed. Lewis's reply, as I have indicated, is to maintain that we should infer from the truth of the conditional that the intuitive standards of similarity are not relevant. We should recognize that the appropriate standard of similarity will include something like the following ranking of how important various forms of differences are: first (most substantial), the existence of many miracles (violations of our laws), second, the absence of an exact matching of particular facts over large regions of space time, and third, the occurrence of a small number of miracles.

Now these criteria of similarity may well engender the right result in each case. However, it seems to me problematic that they have no pretheoretical plausibility and are derived solely from the need to make certain conditionals come out true and others false. For it is now quite mysterious *why* we should have evolved such a baroque notion of counterfactual dependence. Why did we not, for example, base our concept of counterfactual dependence on our ordinary notion of similarity? As long as we lack answers to these questions, it will seem extraordinary that we should have any use for the idea of counterfactual dependence, given Lewis's description of it; and so that account of our conception of the counterfactual conditional must seem psychologically unrealistic.

Finally, let us examine some further difficulties that arise when the a priori component of Lewis's theory is supplemented by the addition of his vital a posteriori hypothesis.

6. *Oversophistication.* The predominantly future orientation of counterfactual dependence, and causation, fall out as consequences of Lewis's theory only relative to a contingent, empirical assumption regarding the asymmetry of overdetermination. He assumes that given a hypothetical change in the actual course of events, it would require many miracles to preserve the actual future, but it would be relatively easy to reconcile that hypothetical change with the actual history of the world. True, some miracle would have to be supposed (assuming determinism) in order to preserve the past, but not on the scale of what would be needed to perfectly shield the future from that change.

It may seem, contrary to this assertion, that many contexts may be found in which it would be just as easy to shield the future from a



hypothetical change as to shield the past. Consider, for example, the counterfactual conditional

If his chair had been one foot to the left at 3 pm, then the rock would have hit him

The antecedent may be reconciled with our actual history before 3 pm by imagining a miraculous sudden jump in the chair's position just before 3 pm. But can we not similarly square the supposition with our actual future by imagining a miraculous sudden jump by the chair back to its original position just after 3 pm? No, says Lewis. Such a jump would not do the trick, for the chair at 3 pm in its hypothetical, temporary position emitted light waves and gravitational forces that are not exactly like the waves and forces it would have emitted if it had not been there then. Therefore, to obtain an exact match with the future, we need to imagine not only the chair jumping back but also many further miracles in order to transform the waves and forces emitted from one position into waves and forces that seem to have come from another position. But the presence of so many miracles would make for a world that is very unlike our own. That is why, if the present were different, the future would have to be different.

My quarrel with this strategy is that it is too scientifically sophisticated. We have presumably been using counterfactuals for thousands of years and have always regarded the future as counterfactually dependent on the past. It cannot be that the ground for such a view lies in the province of contemporary physics. If it were, as Lewis (1987) contends, a matter of plain observational fact that the future grossly overdetermines every event, then it would be legitimate to employ an awareness of that fact in the explanation of our linguistic behavior. But it seems to me that as things are, the fact (if, indeed, it is a fact) is fairly inaccessible—unknown to most people, even today, let alone to our ancestors. Consequently the evaluation of counterfactual conditionals cannot be conducted on the basis of such knowledge.

Lewis does attempt to provide support for his contention, but the argument is not convincing. He points out that detective stories written for the general public presuppose that crimes leave traces. However, it seems clear, in the first place, that we do not take for granted that a 'perfect' crime is impossible (although, not surprisingly, such an event would not be good material for a detective story). And, in the second place, even if clues are presented that do point unambiguously to the criminal, it is not generally supposed that his identity must be *over*determined by the clues.

7. *Empirical implausibility.* Moreover it is not at all obvious that Lewis's empirical assumption is even correct, let alone common knowledge. No doubt there is at least a grain of truth in it, provided by the fork asymmetry: the fact that correlated events have characteristic common causes but not always a typical common effect. But what Lewis needs is a very extreme version of this phenomenon. He must assume that *every* event is one of the later endpoints of a normal fork. This is not merely the trivial claim that every event has 'siblings' (i.e., other events with the same cause). It claims, in addition, that the common cause is determined by, and may be inferred from, each of the effects on its own. Lewis does not, however, give grounds for a thesis of such generality, and I see no reason to accept it.

8. *Backward causation.* Causal overdetermination is, as we noted earlier, the production of an event by more than one causal chain, each of which would have been sufficient on its own for that outcome. Now, according to Lewis, it is the nomological overdetermination of the present by the future that leads to the conclusion that if the present were different, the future would be too, which leads in turn to the future direction of causation. This idea, however, has the following counterintuitive consequence. Consider an event that happens to be very heavily causally overdetermined: for example, a collision caused by several particles simultaneously reaching the same point in space. If that collision had not occurred, then the course of history leading up to it may nevertheless have been as it actually was, but only provided that numerous miracles occur to prevent each of the particles from arriving at that spot when it did. But, as Lewis has argued in connection with the future consequences of an event, this is too high a price to pay. A closer possible world is one in which the miracles are not needed, since the recent history of the world is different and does not involve those particles moving in that way. Thus we have a past-oriented counterfactual and therefore a case of backward causation. But this is not a welcome result. Surely not every case we would normally describe as substantial causal overdetermination is really a case of backward causation!

9. *Jackson's modification.* Frank Jackson (1977) has developed a theory of counterfactuals that combines elements from both of the two major strategies we have discussed. His view is Goodmanian insofar as the truth of a counterfactual is said to depend on whether the truth of its consequent is determined through laws of nature by conditions including the truth of the antecedent. But it borrows from Lewis the way of specifying exactly what the determining condition



must be: namely, a world history just like actuality, leading up to a state of affairs containing the truth of the antecedent and otherwise as similar as possible to the actual state of the world at that time. In other words, to verify a counterfactual conditional of the form, 'If $p$ were true (at time $Ta$), then $q$ would be true (at time $Tc$),' simply establish that the history of the world prior to $Ta$ and the actual state of the world at $Ta$—changed minimally so as to accommodate the truth of $p$—would determine, given laws of nature, the truth of $q$ at $Tc$. Bennett (1984) has proposed a similar account, except that in his view the determining condition does not include the history of the world prior to $Ta$ but consists solely in the minimally altered state at $Ta$.

Not surprisingly, this sort of idea inherits some of the difficulties of the views from which it was derived. Goodman's approach faltered on a circularity that came from supposing that the determining condition would have to be restricted to facts that would continue to obtain even if the antecedent were true. In Jackson's and Bennett's theories this problem is handled by allowing that the determining condition has to be reconciled with the truth of the antecedent, though in the most 'economical' possible way. This is tantamount to accepting Lewis's theory for just those counterfactuals whose antecedents and consequents belong to the same time slice. Therefore many of our objections to Lewis's theory will apply. Indeed, Jackson's own concluding criticism of Lewis may be turned against his own view (I have inserted the material in square brackets):

> Some similarities [to the actual world] between $Ta$ and $Tc$ [and *during* $Ta$] are important and some are not, and which are and which are not, should be part of the *output* of a theory of counterfactuals and not part of its input. Instead of decisions about which similarities in particular fact after [and during] the antecedent-time are to be preserved determining which counterfactuals with that antecedent are true, it is the other way around: the true sequential counterfactuals with that antecedent settle the similarities preserved. Hence a theory of sequential counterfactuals ought to *yield* the subsequent [and simultaneous] similarities, not draw on them the way Lewis's [and Jackson's] does. (1977, p. 8)

Moreover accounts that restrict the determining condition to occurrences no later than $Ta$ are faced with a problem of their own. Consider a *random* event (e.g., the toss of a coin yielding heads) that takes place after $Ta$. Intuitively such an event belongs among the circumstances in which the antecedent is imagined to hold, so that it can be

true to say, "If I had guessed heads, I would have been right". Bennett tries to handle such cases with an ad hoc modification of his analysis. But there is no natural way to obtain results of this kind within the spirit of the views just described.

This completes my discussion of Lewis's program. His account of causation began with a critical appraisal of its main competition, namely, the sort of regularity theory that I have been advocating here:

> It remains to be seen whether any regularity analysis can succeed in distinguishing genuine causes from effects, epiphenomena, and pre-empted potential causes—and whether it can succeed without falling victim to worse problems, without piling on the epicycles and without departing from the fundamental idea that causation is instantiation of regularities. I have no proof that regularity analyses are beyond repair, nor any space to review the repairs that have been tried. Suffice it to say that the prospects look dark. I think it is time to try something else. (1973b, p. 557)

It seems to me that the pendulm has swung. What Lewis said about regularity analyses is now a fair assessment of the counterfactual approach.

# 11
# Decision

## 1. *Decision theory in light of Newcomb's problem*

You should act in some way when desirable events are to be expected if you do. Or so one might suppose, but for the following problem. The expectations of an action divide into two sorts: confidence that it would *cause* desirable events and belief that it would be highly *symptomatic* of their prior occurrence. And it may appear that only in the former case does one have a decent motive: only the causal implications matter; the merely evidential implications of your act are irrelevant to its choiceworthiness. So it seems that the initial (evidential) principle should be amended: do something if you expect it would *bring about* desirable *results*. This causal point of view has lately become orthodox opinion among philosophers of decision theory. However, I disagree with it, and my primary aim in this chapter is to explain why, and to argue in favor of the purely evidential conception of rational choice.

A second, intimately related, aim is to explain the time asymmetry of rational choice. Why do we act for the sake of desirable future events but not for the sake of the past? Obviously, if the causal theory of decision is correct, then the temporal orientation of rational choice follows immediately from the direction of causation. We act for the sake of the future because we act for the sake of events that we can cause, and those can only be in the future. However, if the evidential conception of rational choice is correct, and I think it is, then a different account of the decision time asymmetry must be provided. Given *that* conception of rational choice, the time asymmetry could exist only if past events fail to stand in the appropriate evidential relation to our decisions. We shall see, later on, how this is so.

The evidential principle, in one of its precise forms, is the requirement to maximize expected desirability. In other words, for every act under consideration, multiply the desirability of each alternative eventuality by its probability relative to the act in question, and add these products together, thus obtaining the act's expected desirability;