CHAPTER 8

COUNTERFACTUALS

JULIAN REISS

8.1. Introduction: Varieties of Counterfactuals in the Social Sciences

Counterfactuals are conditional or "if-then" statements that describe what *would* follow if something *were* the case. Using A for its antecedent, C for its consequent, and > for the counterfactual conditional, a counterfactual statement has the general form "A > C" or "Had A been, C would have been."

Statements of that form are used in various functions throughout the social sciences. Consider the following quotations:

> The judgment that, if a single historical fact is conceived of as absent from or modified in a complex of historical conditions, it *would* condition a course of historical events in a way which would be different in certain *historically important* respects, seems to be of considerable value for the determination of the "historical significance" of those facts. (Weber 1949 [1905], 166; emphasis original)
>
> The counterfactual approach to causal analysis for this problem focuses on the collection of potential responses Y:= ($Y_i(u)$: $i \in T$, $u \in U$), where $Y_i(u)$ is intended to denote "the response that would be observed if treatment i were assigned to unit u." (Dawid 2000, 409)
>
> In the field of evaluation, Mohr (1995) points out that the use of the counterfactual is essential in impact evaluation, as it provides the alternative against which the program's impact can be measured. In its broadest sense, the counterfactual is an estimate (either quantitatively or qualitatively) of the circumstances that would have prevailed had a policy or program not been introduced. (Cummings 2006)
>
> But what if we envision, as we have done, a world without medicine? There would be no primary care, nor would there be surgery, nor other specialties, nor pharmaceuticals. The medical model as well would disappear. At first glance, the effect would seem to be devastating. Surely more people would die and die before their time. Our thought experiment reveals a different picture. (Markle and McCrea 2008, 129–30)

> One implication of this analysis is that the value of the NAIRU concept depends on the monetary regime. If we lived in a world where inflation was close to white noise, rather than highly persistent, then adaptive expectations would be a bad approximation to optimal behavior. (Ball and Mankiw 2002, 115–136)

In the first three quotations, the counterfactuals are intimately linked with causality. The first is an example of a long tradition of establishing actual causation by means of the so-called but-for test, which is prominent in history and the law. According to this tradition the legal scholar can test whether some action was a cause of harm by asking whether the harm occurred but for the action or "would the harm have occurred if the action hadn't?". Similarly, a historian can determine if a decision of a historical actor was a cause of an outcome of interest by asking "did the outcome occur but for the decision?".

The second quotation is a description of the potential-outcomes framework in statistics. Its core quantity is the so-called individual causal effect (ICE), defined as:

$$\mathrm{ICE}(u) = Y_t(u) - Y_c(u),$$

which is the difference between the outcome that would have obtained had the unit been treated (i.e., in the treatment group t) and the outcome that would have obtained had the unit not been treated (i.e., in the control group c). The approach originated in statistics (e.g., Rubin 1974, 1977; Holland 1986) but is now widely applied throughout the social sciences (Morgan and Winship 2007; Morton and Williams 2010; Heckman 2005).

The third case is closely related. Here counterfactual worlds are constructed in order to evaluate the impact of policies. Questions are asked such as "What would the outcome be if our policy had been implemented?" or "What if it had not been implemented?" (For a discussion, see Reiss and Cartwright 2004; Cartwright 2007.)

The fourth quotation stems from a book written by two medical sociologists who imagine a world without medicine. Such counterfactual speculations are done for a variety of cognitive purposes—that is, not only to estimate causal effects—and can be found in many disciplines inside and outside of science. In world history the genre of "virtual history" has become popular in recent years (Ferguson 1997; Tetlock and Belkin 1996; Tetlock et al. 2006; Hawthorn 1991). The related genre of "alternate history" is a popular branch of fiction (see, for instance, Hellekson 2001).

Finally, the fifth quote contains in fact two counterfactuals. The explicit counterfactual ("if inflation was white noise, adaptive expectations would not be a good approximation to optimal behavior") concerns the justification of an assumption of an economic model. Such model-based counterfactual reasoning is ubiquitous in branches of the social sciences that are heavily mathematized such as theoretical economics and political science. The implicit counterfactual is part of the concept of a NAIRU or non-accelerating inflation rate of unemployment. This is the rate of unemployment that *would* obtain *were* inflation non-accelerating. Counterfactually defined concepts are frequent in economics.

In this chapter I will only be concerned with counterfactuals insofar as they relate to causal inference about singular events; that is, with the first type and type three to the extent that causal inference is the purpose of the counterfactual speculation. As mentioned above, this type of counterfactual is specifically relevant to causal analysis in the historical sciences and in the law.

8.2. Starting Points

Counterfactual speculation has not always had a good press. In the words of Marxist historian Edward Carr, it is a mere parlor game, and he recommends: "Let us get rid of this red herring once and for all" (Carr 1961: 91f.; but see also Fischer 1970; Thompson 1978). To some extent, this critical attitude has to be understood as being a product of positivism. If, as positivists held, only that which is observable is meaningful, then counterfactuals should have no place in science or indeed anywhere in human reasoning. But counterfactuals have stood the test of time with more success than positivism, and a world without at least some speculation about what would, could, or might have been would be utterly impoverished. Indeed, many cognitive psychologists today believe counterfactual speculation is at the heart of learning about the world (Gopnik 2009; Sloman 2005).

In many cases, we assert counterfactuals with as much confidence as factual claims. Had I just dropped my coffee mug, it would have fallen to the ground. If I hadn't written this chapter, you would not read it now. In other cases, it is similarly clear that a counterfactual is *not* assertable: "Had I rolled a die, it would have landed on 6 (or on any other specific number)." Or suppose that Jones is an average golf player. On this occasion he slices the ball but, as it happens, it hits a tree, which deflects the ball straight into the hole. In this case, we would not be justified in asserting that had Jones not sliced the ball, he would still have made a hole-in-one. But to assert the opposite would be wrong too. We just don't know.

Arguably, many historical counterfactuals are more of the latter type than of the former. Had the 2008 financial crisis occurred if the US government had not allowed Lehman Brothers to fail? Would European values be as dominant in the world as they are had Themistocles lost the battle at Salamis in 480 BC? Counterfactuals such as these are thorny.

A presupposition I make in this chapter is that counterfactual speculation is not idle, at least not always. That is, at least sometimes, the relevant counterfactuals are of the former, evaluable, and not of the latter, inscrutable, type. In other cases we may learn something useful while trying to establish a specific counterfactual even if it turns out that we cannot know its truth value with a reasonable degree of confidence.

As mentioned in the introduction, this chapter focuses on counterfactuals that are closely related to claims about actual causation. Some philosophers, most

notably David Lewis, have maintained that counterfactual dependence is all there is to actual causation, and that actual causation, in turn, just *is* causation.

I deny both of these ideas. That there is more to actual causation than counterfactual dependence is evident when one considers cases of so-called redundant causation where two or more causes compete in their bringing about an effect. To give an example, suppose that two campers in different parts of the woods leave their campfires unattended. A forest fire results, but with or without either campfire (though not without both). As regards the law, both campers' actions are equally causes of the fire, even though the fire counterfactually depends on neither.

Moreover, actual causation does not exhaust the kinds of causal relations that exist and that social scientists are interested in. Factors can be causally relevant to an outcome or a type of outcome without being the actual cause of it. In particular, social scientists are often interested in generative causal mechanisms, which often do not stand in relations of actual causation to their effects (see, for instance, Goldthorpe 2001; for a critical discussion, see Reiss 2007).

Nevertheless, claims about actual causation are important in the social sciences and the counterfactual approach to actual causation is a significant one, even if it is not universally valid. Therefore, rather than dismissing any attempt at developing a counterfactual account of actual causation upfront, I will here take it seriously and address some more specific questions such as:

- How precisely do counterfactuals and causation relate?
- How can we use knowledge about counterfactual dependencies for causal inference?
- How do we support claims about counterfactual dependence with evidence?

In what follows I will first introduce a philosopher's answer to the first question and examine whether that account doubles up as an answer to the second. After pointing out severe difficulties with both, I will move on to a group of social scientists who have developed an account to address the second and third issues. It will also turn out to be flawed. I will then introduce and discuss a third approach regarding the relationship between counterfactuals and causation, argue that it is the most convincing but point out that eventually it shows that this way of thinking about the relationship between counterfactuals and causation implies that the second and third are hard nuts to crack indeed.

8.3. The Philosophers' Approach

A brief quotation from David Hume's *Enquiry Concerning Human Understanding* nicely summarizes two ideas concerning causation that have dominated the philosophical debate in the twentieth century (Hume 1777 [1902], section 7; emphasis original):

> *we may define a cause to be an object followed by another, and were all the objects, similar to the first, are followed by objects similar to the second.* Or, in other words, *where, if the first object had not been, the second never had existed.*

These two definitions of cause, though equivalent to Hume, refer to the regularity and the counterfactual accounts, respectively. The regularity account, dominant in analytical philosophy through the 1960s, asserts that causation is a form of constant conjunction: For a factor to cause another means to be universally associated with it, among other things.[1]

David Lewis's account of causation, which has been hugely influential after the demise of the regularity theory, builds on Hume's second definition. Specifically, it lays out two sufficient conditions for causation:

C causes *E* if:

- *C*, *E* are actual, distinct events; and
- $\neg C > \neg E$.

The first condition is to rule out certain counterexamples that arise from logical, conceptual, and other connections that induce counterfactual dependence for non-causal reasons. Thus, the appearance of the evening star counterfactually depends on the appearance of the morning star but the former event is not a cause of the latter. The second condition is the but-for test for causality: We judge an event to be the cause of another if the second would not have occurred but for the first.

To evaluate the counterfactual $\neg C > \neg E$ Lewis invokes possible-worlds semantics. He first stipulates that all possible worlds can be (weakly) ordered in terms of distance to the actual world. For him, one world is closer to the actual world than another if the first is more similar overall to the actual world than the second. Lewis therefore assumes that different aspects of similarity trade off against each other: When Jill is more similar than Beth to Mary in terms of height and Beth is more similar than Jill in terms of weight, there is a sense in which Jill is more similar to Mary than Beth overall (for instance, because height counts more for overall similarity than weight).

Second, Lewis defines the counterfactual $\neg C > \neg E$ to be non-vacuously[2] true if and only if some $\neg C$-world in which $\neg E$ holds is closer to the actual world than any $\neg C$-world in which *E* holds. In other words, in order to evaluate whether an actual event *C* causes another, distinct event *E*, we have to ask whether there is any possible world in which *E* obtains even though *C* does not that is more similar to the actual world than the most similar world in which neither *C* nor *E* obtain; if such a world exists, *C* does not cause *E* but if it does not exist, *C* causes *E*.

In response to counterexamples (e.g., Fine 1975) Lewis later revised his theory, rendering it more precise thereby. In particular, he proposed the following system of weights or priorities for judging similarity (Lewis 1986, 47):

1. It is of the first importance to avoid big, widespread, diverse violations of law.
2. It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.

3. It is of the third importance to avoid even small, localized, simple violations of law.
4. It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

There remain some counterexamples but we can ignore these here.[3] Assume therefore for the sake of the argument that Lewis's theory perfectly tracks our ordinary concept of counterfactual dependence. My question here is whether we can employ his ideas about counterfactual dependence for causal inference in the social sciences.

From the point of view of the social scientist, the problems with Lewis's account are twofold. There is, first, a semantic problem. In Lewis's semantics a putative cause is removed from world history by inserting a miracle—a violation of a natural law—just before the cause occurs. The departure from actuality is minimal in *one* sense: A minimal number of natural laws is to be violated (relative to actuality) in order to realize the counterfactual antecedent. But it is not minimal in a different sense: It contradicts a law claim we hold to be true of the world. Many social scientists prefer to evaluate counterfactuals in a way that does not contradict firm beliefs they hold about the world. Consider the following criticism of the so-called early-warning counterfactual concerning the Cuba missile crisis made by political scientists Richard Ned Lebow and Janice Gross Stein. The early-warning counterfactual asserts that if only Kennedy had issued a timely warning in the spring of 1962, Khrushchev would not have sent missiles to Cuba.[4] Lebow and Stein argue (Lebow and Stein 1996, 129):

> In April, before the conventionals buildup began, Kennedy had no reason to suspect a missile deployment, and months away from an election campaign, had no strong political incentive to issue a warning. To sustain the early-warning counterfactual, other counterfactuals would have to be introduced to provide foreign or domestic motives for warnings in April.

Under a Lewisian reading of counterfactuals, such a criticism would be beside the point. In the deterministic world Lewis conceives *any* counterfactual antecedent has to be brought about by miracle, by violation of natural law. Fewer, more localized violations of law constitute a smaller departure from actuality to be sure but in principle there are no differences among the laws. By contrast, many social scientists make such distinctions. Certain events seem more haphazard than others and can therefore more confidently be removed from the course of history. The assassination of Archduke Franz Ferdinand is often considered such a haphazard event (Lebow 2010, 44):

> I use realistic here in a more subjective and psychological sense of not violating our understanding of what was technologically, culturally, temporally, or otherwise possible. In chapter 3, I imagine a world in which Archduke Franz Ferdinand and his wife, Countess Sophie, returned alive from their visit to Sarajevo. This counterfactual is eminently plausible because their assassination was such a near thing, and never would have happened if the archduke and those responsible for his security had acted sensibly either before the first, unsuccessful attempt on his life or in its immediate aftermath.

That Kennedy did not issue an early warning is less of an accident of history. Many things would have to have been different for Kennedy to have issued an early warning: He would have to have had better intelligence, which would have required institutional differences. Or he would have to have stood in the need of a national political success, which also would have required a different institutional setting.[5]

Not all Lewisian counterfactuals are therefore regarded as admissible by social scientists. As we will see, there is a further difference that requires a more radical departure from Lewis's semantics.

The second problem is epistemic. Lewis parted with the Humean tradition in important ways. Hume sought to reduce the concept of cause to constant conjunction because, being an empiricist, he thought that causation is suspicious qua not being observable. We can observe one billiard ball moving towards another and, upon impact, the second ball moving, but we cannot observe the "push," the causal power of the first ball to move the second. Regularities are, however, straightforwardly observable and therefore an analysis of cause in terms of regularity makes sense from an epistemic point of view: An epistemically inaccessible (because unobservable) concept has been analyzed in terms of an epistemically accessible (because observable) one.

This is not so on the Lewisian analysis. Possible worlds, laws, and miracles are certainly no more and quite possibly much less epistemically accessible than the concept of cause we are analyzing. But then trying to use the account for causal inference would be futile.

This is easy to see. In Lewis's account the notion of natural law plays a crucial role. But laws—in the sense of strict regularities—are few and far between in the social world. Few events, even those we believe to have explained causally, will fall under a natural law. At any rate, such laws, to the extent that they exist, are not known by social scientists and can therefore not be used for causal inference via counterfactuals.

8.4. The Social Scientists' Approach

Social scientists working in the counterfactual tradition tend not to develop semantics for counterfactuals as such but rather a list of desiderata counterfactuals should realize in order to be regarded as admissible or "good." The following is a typical list.

8.4.1. Specificity

In Lewis's examples and those of his followers, there is usually a reasonably unambiguous way to remove the putative cause-event from the course of history.[6] Billy's not throwing rocks at a bottle (that would have shattered it) means that Billy is standing still rather than throwing a hand grenade (the example is due to Lewis

2000). The dog's not biting off the assassin's right forefinger (who subsequently has to use his left forefinger to detonate the bomb) means that the dog is doing nothing to the assassin rather than biting in his throat and thereby killing him (the example is due to McDermott 1995). Usually, it is therefore unnecessary to say more about the counterfactual antecedent apart from it being the proposition that some actual event did not happen.[7] Social scientists have to provide considerably more detail. According to Lebow and Stein the antecedent of the Cuba crisis counterfactual "Had the United States attacked the missile bases, the Soviet Union would have responded to an attack on Cuba with military action of its own" is too unspecific to entail an unambiguous counterfactual consequent: The Soviet response would presumably have depended on whether the attack had been a surgical air strike or an invasion that would have toppled the Castro government (Lebow and Stein 1996, 139). On the other hand, the antecedent should not be overdescribed because outcomes are usually not dependent on other events in all their historical detail. Thus, while the Soviet response depends on whether the US attack would have been an air strike or an invasion, it does not depend on the exact timing of the attack, on precisely how many planes or ships would have been involved, and on who is commanding them.[8]

8.4.2. Cotenability

Intuitively, the cotenability desideratum says whatever else we assume in order to make the counterfactual true should not be undermined by the counterfactual antecedent.[9] That is, it should not be the case that the researcher assumes some statement *B* to be true in order to make the judgment regarding *C* but if the counterfactual antecedent *A* were true, *B* would be false. To give a classical philosopher's example—because it nicely exhibits the difference between the philosophers' and the social scientists' approach—consider the following (Lewis 1986, 33):

> Jim and Jack quarreled yesterday, and Jack is still hopping mad. We conclude that if Jim asked Jack for help today, Jack would not help him.

In this case, *A* is "Jim asks Jack for help today," *B* is "Jim and Jack quarrelled yesterday," *C* is "Jack does not help Jim." In order to derive the consequent from the antecedent, use of the additional clause *B* is made. The next two sentences show that *A* and *B* are not cotenable (Lewis 1986, 33):

> But wait: Jim is a prideful fellow. He never would ask for help after such a quarrel; [. . .]

Thus, if it were the case that Jim asked Jack for help today, it would have to have been that Jim and Jack did not quarrel. The passage continues (Lewis 1986, 33):

> In that case Jack would be his usual generous self. So if Jim asked Jack for help today, Jack would help him after all.

If it is known that Jim is a prideful fellow, *A* and *B* are not cotenable: *A* counterfactually entails the negation of *B*. Therefore, *B* cannot be used in deriving the consequent "Jack does not help Jim."

Jon Elster criticizes Robert Fogel's work on the American railroads (Fogel 1964) on these grounds (Elster 1978). In Fogel's counterfactual the antecedent *A* is "America has no railroad," the auxiliary *B* is "the combustion engine is invented earlier," and the consequent *C*, "the US social product is about the same as the actual." Elster effectively argues that *A* and *B* are not cotenable because under any theory of technological innovation, if there had been no railroad, there would have been no combustion engine either. The theory of technological innovation thus plays the same role as Jim's character trait "being a prideful fellow": Given this principle, *A* counterfactually entails the negation of *B*.

Lewis, by contrast, goes on to argue that the counterfactual "Had Jim asked Jack for help today, Jack would not help him" is true after all. This is because, Lewis argues, (a) counterfactuals are vague and (b) we ordinarily resolve the vagueness in such a way that counterfactual dependence is asymmetric: The future counterfactually depends on the present and past but not vice versa (ibid., 34). Accordingly, the counterfactual "If Jim had asked Jack for a favor today, there would have to have been no quarrel yesterday" is false under the "ordinary resolution of vagueness." Lewis does not require cotenability because antecedents, implemented by miracle, are cotenable with any other truth. This way Lewis avoids backtracking counterfactuals. In the present case *A* and *B* are cotenable because even if there was a quarrel yesterday Jim would have asked Jack for a favor today because in Lewis's semantics the law that says that prideful fellows do not ask for favors a day after a quarrel is broken. Hence it does *not* follow that if there had been a quarrel, Jim would not have asked Jack for a favor.

Social scientists, by contrast, aim to keep as much as possible about historical actors' situations and dispositions intact (see section 8.4.3). In order to achieve cotenability, then, counterfactuals will sometimes have to backtrack. Whether they do so would depend on the nature of the involved events and generalizations as well as the strength of the evidence in their favor. In our case the relevant events and generalizations are: "There was a quarrel between Jim and Jack yesterday," "Jim is a prideful fellow," and "prideful fellows do not ask for favors a day after a quarrel." If the quarrel had been accidental and unimportant—Jim and Jack are good friends; they both have amiable characters; nothing important for understanding their lives happened to caused the quarrel; what was responsible was an accidental splashing of Jack by a careless driver and Jack's resulting foul mood—and at the same time there is good reason to believe that Jim is a prideful fellow and that prideful fellows don't ask for favors a day after a quarrel, in order to implement the antecedent, one would have to remove the quarrel and consequently judge the counterfactual "Had Jim asked Jack for a favor today, Jack would oblige" to be true. This is a backtracking counterfactual as "Had Jim asked Jack for a favor today, there would have to have been no quarrel" would also be true.

If, by contrast, the quarrel was a major event in their lives and nonaccidental—for instance, caused by grievances both friends have accumulated over the years—then the antecedent could only be true if Jim isn't (believed to be) such a prideful fellow after all or the generalization about prideful fellows is (believed to be) unreliable. In this case, the counterfactual "If Jim had asked Jack for a favor today, Jack would not oblige (because he is still hopping mad)" would be judged to be true. If both (a) the quarrel was important and (b) belief in Jim's character and the generalization is strong, the antecedent would not be cotenable with auxiliary beliefs.

In the already mentioned paper on the Cuba crisis by Lebow and Stein one finds an example for background information that is not cotenable with a proposed antecedent. Lebow and Stein argue that the counterfactual "Had President Kennedy issued a timely warning in the spring of 1962, Khrushchev might not have sent missiles to Cuba" does not satisfy the cotenability desideratum because for Kennedy to issue a warning there would have to have been an election campaign or different intelligence. Those in turn would require further changes in the past of the antecedent event. It would have been very unlikely that these changes would or could have happened. Therefore, the only way to make the antecedent cotenable with certain background beliefs much history would have to be rewritten. But doing so is proscribed by the following desideratum.

8.4.3. Historical Consistency

This is an interesting desideratum because it sounds similar to one of Lewis's criteria (see above: "(2) maximize the spatio-temporal region throughout which perfect match of particular fact prevails") but is in fact different in crucial ways.[10] Tetlock and Belkin explain that the substance of this desideratum is that possible worlds should (Tetlock and Belkin 1996, 23):

> (a) start with the real world as it was otherwise known before asserting the counterfactual; (b) not require us to unwind the past and rewrite long stretches of history; (c) not unduly disturb what we otherwise know about the original actors and their beliefs and goals.

(a) looks like the requirement that the *closest* possible world is the relevant one, (b) looks like the prohibition of backtracking counterfactuals, and (c) looks like the requirement to avoid big, widespread violations of laws. But the interpretations of (b) and (c) are in fact quite different.

I already discussed that social science counterfactuals sometimes involve backtracking. Here is an example of how aiming to satisfy the desideratum of historical consistency can lead to a backtracking counterfactual. Yuen Foong Khong 1996 asks if World War II could have been avoided if the UK foreign policy had been more confrontational. A Lewis counterfactual would make the antecedent true by miracle: by a surgical intervention that changes nothing but the UK foreign policy. In that possible world the UK would still be led by Neville Chamberlain and his cabinet, only their policies would be different. But this would violate what we know about the UK leaders at the time and therefore the desideratum of historical consistency. We know that Chamberlain was averse to policies that would have risked war because the horrors of the World War I were still in his memory and that of the British public, because he felt that Britain was militarily ill prepared, and because he had

a firm belief in himself, his diplomacy, and Hitler (Khong 1996, 100–1). A confrontational Britain with Chamberlain's cabinet in the saddle is therefore historically inconsistent. However, backtracking ever so slightly allows us to implement the antecedent after all. Because more confrontational potential prime ministers (in particular Winston Churchill, Anthony Eden, and Duff Cooper) were at the top of British politics at the time and they could have been prime ministers given the UK electoral system, a world with a UK that rejects appeasement because a different prime minister heads the government is conceivable.

Hence, there may be a trade-off between Tetlock and Belkin's desiderata (b) and (c). Unless an event is a particularly close call—such as the failed assassination of Ronald Reagan or the successful assassination of Archduke Franz Ferdinand—undoing it will either involve more than one counterfactual (say, Britain takes a more confrontational stance in 1938 *and* Chamberlain has different political convictions) or require backtracking. The main difference to the philosophers' semantics is therefore that the proscription of backtracking is not absolute. Backtracking is to be avoided but not at the cost of falsifying what we otherwise know about historical contexts and actors' motivations and beliefs.

There is another, related difference. As mentioned above, in Lewis's semantics every violation of a natural law is on par. Therefore, in order to measure the distance between possible worlds we merely have to count the number of violations. For social scientists, "the *nature* of the changes made by the experiment are [*sic*] nevertheless more important than the *number* of changes" (Lebow 2010, 55; emphasis in original). In a deterministic world, there is no difference between moving Hinckley's bullet from Reagan's lungs to his heart and changing one of Kennedy's decisions in the Cuba crisis. Both require a violation of laws. But from a historian or social scientist's point of view, there are massive differences. That Hinckley's bullet punctured Reagan's lung but not his heart was an accident, and a failure from the point of view of Hinckley's intentions. Kennedy's decisions were not accidents, or at least are not regarded as such, but rather the outcome of deliberative processes that make use of Kennedy's beliefs and motives and those of his aides. It requires a small miracle to change the trajectory of Hinckley's bullet. But it requires a far larger miracle to make Kennedy show greater resolve given he had neither domestic nor foreign reasons to do so.

Norms, not natural laws, play a role in determining the importance of changes (cf. Hitchcock and Knobe 2009). In the case of Franz Ferdinand's assassination, for example, it was only the violation of norms that enabled the attack: He had been warned of possible assassination attempts and could easily have returned to Belgrade immediately; security was unusually bad; his touring car took a wrong turn; a failed assassination attempt preceded the successful one and could have made the archduke more cautious. We make the counterfactual true by imagining that Franz Ferdinand complied with behavioral norms and norms of prudence. By contrast, in order to make the Kennedy counterfactual true, we would have to imagine Kennedy to violate norms (such as the norm that a US president should not issue a warning unless he has good reason to do so). It is therefore that the latter counterfactual requires more rewriting than the former.

8.4.4. Consistency with Well-established Theoretical Statistical Generalizations

Some social scientists think there are no accepted theories in (some of) the social sciences at all.[11] Others, most notably rational-choice theorists, structural realists,[12] and others of nomothetic bent, think that all social events fall under a theory. But whatever our take on that issue, counterfactual speculation is impossible without generalizations of one sort or another. That Reagan would have died had Hinckley's bullet taken a minimally different trajectory is only true only to the extent that humans normally die when their heart is punctured and they can't be given immediate surgery. Even particularists such as Lebow and Stein make generalizations of that kind. In their book on the cold war (Lebow and Stein 1996), for instance, they draw on a psychological theory of decision-making under stress to derive the consequent that if Kennedy had shown greater resolve, Khrushchev would still have deployed missiles.

Importantly, no matter how much we are inclined to think that theories have limited and local validity at best, they must satisfy minimal projectability requirements. Insofar as our counterfactuals are to be based on evidence, the theory has to cover at least two cases: that for which we have direct evidence and from which we are projecting and the counterfactual case we are projecting on.

To give assertability conditions analogously to Lewis's semantics one could proceed as follows (cf. Reiss 2008). Let H be a historian or social scientist with background beliefs about relevant evidence and causal generalizations B, and X a historical context such that $\neg A$. Then, relative to B, the counterfactual $\neg A > \neg C$ is assertable iff

a. A, C actually obtained;
b. $\neg A$ is historically consistent and specific enough to allow H to make a judgement regarding $\neg C$;
c. $\neg A$, $\neg C$ are cotenable and consistent with well-established theoretical and statistical generalizations;
d. H judges $\neg C$ to obtain in X.

The social science desiderata are very plausible and useful but hardly rigorous. The first desideratum does not tell us precisely how specific the counterfactual antecedent is to be described. The third uses vague terms such as "long" stretches of history and "unduly" disturb what we know, but how long is long and when do we unduly disturb what we know about history? Moreover, if I am right in saying that there is sometimes a trade-off between avoiding falsifying our historical knowledge and avoiding backtracking, how are we to trade off these desiderata? The fourth may be too weak or too strong, depending on the reading of "consistency," and it is vague and ambiguous. If by consistency logical consistency is meant, it is very weak, especially if there are few well-established generalizations. Suppose Lebow and Stein are right in denying that there are any well-established social science theories. If so, nearly every counterfactual will be true as long as it does not

contradict the laws of physics or other natural sciences. On the stronger and more plausible (albeit literally incorrect) reading, the desideratum demands that the counterfactual follow deductively from the conjunction of antecedent, auxiliary assumptions and the generalizations. In this case, d) of the above assertability conditions above would be superfluous because $\neg C$ would already be entailed by c). However, very few counterfactuals will be true if there are only few social science generalizations. Further, "well-established" is a vague and ambiguous term. Economists will regard rational-choice theory as well-established but many other social scientists think it is false. Established according to whose criteria? How much evidence do we need in order to regard it as well-established?

In order to avoid problems such as these, in the following section I will introduce a (philosophical) theory of counterfactuals that makes use of causal modeling tools. As we will see, the theory preserves the main intuitions of the social scientists' approach and it has the additional advantage of being more precise. At the same time, it makes counterfactuals true relative to a model so that many of the above-mentioned worries regarding the vagueness and ambiguity of the criteria are relegated to assessing whether any given model is one that is good or adequate. Nevertheless, the tool of causal modeling provides a neat language within which one can address these problems.

8.5. Causal Theories of Counterfactuals

Much of the philosophical tradition regards counterfactuals as being analytically more basic than causation and thus aims to provide a counterfactual analysis of causation. But there are good reasons to think that the relationship goes the other way. We judge counterfactuals on the basis of our causal background knowledge. The conviction that my headache would have gone by now if only I had taken an aspirin stems from my knowledge of the causal power of aspirins to relieve headaches (in conjunction perhaps with the more specific causal facts that I am not allergic to aspirin and that they have been effective in me before). This point is stressed by Jon Elster when he charges David Lewis's theory with circularity (Elster 1978, 218):

> My objections to Lewis's theory have been of two kinds. In the first place I have argued, quite generally and without reference to historical counterfactuals, that to explain causality by counterfactuals and counterfactuals by similarity is a circular procedure, as causal importance is an element in our intuitive notions about similarity.

According to Elster, Lewis's theory is circular because judgments of similarity rely on causal judgments (whereas it's not necessarily the case that all causal judgments rely on judgments about counterfactuals or similarity among possible worlds).

The second reason is epistemic. It is controversial whether or not causation is observable. Humeans believe that we can observe only the manifestations of causation in the resulting changes but not the causal *oomph* itself. Others think that causal relations can be observable under certain conditions, and indeed there is some evidence that small children form causal concepts before concepts even of persisting objects (Gopnik et al. 2004). Without taking a stance in this debate, let me make two uncontroversial points. First, counterfactuals are uncontroversially unobservable. Whatever facts there are in the world, there are only facts. A counterfactual refers to a possible but not actual state of affairs and is therefore by its very nature unobservable. Second, there are very reliable, and well understood, methods of causal inference. For many causal claims, we know how to test them. Counterfactuals, by contrast, first have to be translated into a different kind of claim (for instance one regarding laws, similarity among possible worlds, and, indeed, causation) and these then may or may not be testable. Lewis's similarity metric, at any rate, has no empirical counterpart.

There are now various causal accounts of counterfactuals (an early theory is Jackson 1977; more recent accounts include Pearl 2000; Reiss and Cartwright 2004; Maudlin 2007). Here I will introduce and discuss an account by Eric Hiddleston (Hiddleston 2005) because it is both rigorous and has the right semantics for social science applications.

In Hiddleston's theory, a causal model is a triplet <G, E, A>. G is a directed acyclic graph, which consists of a set of variables and arrows or *edges* indicating direct causal relations between some of them. The graph is *directed* in that it does not contain any undirected edges (indicating a correlation or the existence of a common cause). The graph is *acyclic* in that it does not contain any cycles such as X → Y → Z → X. E is a set of structural equations relating the (probabilities of values) of each variable X in G to the values of its direct causes or *parents* pa(X) in G. They represent the causal principles assumed to be true of the modelled scenario. The equations have the general forms:

$$(Y_1 = y_1 \,\&\, \ldots \,\&\, Y_n = y_n) \Rightarrow X = x,$$

$$(Y_1 = y_1 \,\&\, \ldots \,\&\, Y_n = y_n) \Rightarrow p(X = x) = z,$$

for the deterministic and the indeterministic case, respectively, where the Y's are X's parents, the y's their specific values and z is the probability that X has the value x. "⇒" reads "causes." Thus, the entire equations read, "Y_1's having value y_1 in conjunction with . . . and Y_n's having the value y_n causes (the probability of) X to have/having the value x (to be z)".

A is an assignment of values to the variables in G which is possible given E so that no variable X has a value A(X) which the equations say has probability o given the values A assigns to X's parents: For any X, p(A(X) | pa(A(X))) > o.

I will use the appeasement counterfactual as the main example throughout. I will start with an extremely simplified model, which is made more realistic as we go on. Suppose for now that there are only three variables BFP (for British foreign policy), PM (for prime minister) and Cab (for other members of the cabinet). BFP

has the two values A(BFP) = {dove, hawk}, where "dove" means appeasement and "hawk" confrontation, PM has the four values A(PM) = {Chamberlain, Churchill, Cooper, Eden} and Cab has the two values A(Cab) = {dove, hawk}, where the values mean that the majority of other cabinet members prefer appeasement or a confrontational policy, respectively. The actual values are listed first. The equation describing the causal relations among these variables is simply:

PM = Chamberlain & Cab = dove => BFP = dove (BFP = hawk otherwise).

That is, in this model British foreign policy is appeasing if both the prime minister is Chamberlain and the majority of other cabinet members are also doves (and confrontational otherwise).[13] The corresponding graph is described in figure 8.1.

We now need to characterize the concepts of *direct positive influence, positive parents*, and *causal break* (Hiddleston 2005, 640–1). Suppose X is a parent of Y in a model M, X = x, Y = y, Y's other parents are **Z**, and these parents have the values **z**.

Direct positive influence. X = x has direct positive influence on Y = y in M iff

$$p(Y = y \mid X = x \;\&\; \mathbf{Z} = \mathbf{z}) > p(Y = y \mid X \neq x \;\&\; \mathbf{Z} = \mathbf{z}).$$

Y's positive parents in M are $\mathbf{ppa}(Y)_M$ = {X: X = x has direct positive influence on Y = y in M}.

That is, holding fixed Y's other parents, X = x has direct positive influence on Y = y if and only if X's having value x raises the probability of Y's having value y. The positive parents of a variable in a model are all those parents that have a direct positive influence. Both parents in our example are also positive parents.

Causal break. A causal break in model M_i from M is a variable Y such that $A_i(Y) \neq A(Y)$, and for each $X \in \mathbf{ppa}(Y)_M$, $A_i(X) = A(X)$.

Break(M_i, M) = {Y: Y is a causal break in M_i from M}.
Intact(M_i, M) = {Y: $A_i(Y) = A(Y)$ and for each $X \in \mathbf{ppa}(Y)_M$, $A_i(X) = A(X)$}.

A causal break is thus simply a variable that takes a nonactual value in M_i while all of Y's positive parents have their actual values. **Break** is the set of breaks in M_i from M and **Intact** the set of variables for which M_i gives actual values both to Y and its parents. In our example, we can consider a model M_i in which Churchill is prime minister in 1938 instead of Chamberlain. In that model, PM is a causal break, **Break**(M_i, M) = {PM}, **Intact**(M_i, M) = {Cab} and BFP = hawk.

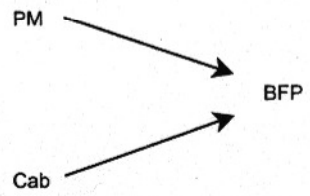

Figure 8.1 Causes of BFP

Two further definitions have to be made. The first one concerns the notion of a minimally altered model. Call a model in which some proposition ϕ is true a ϕ-model. The variable set **Break**(M_i, M) is minimal among ϕ-models if and only if there is no ϕ-model M_k such that **Break**(M_i, M) $\supset$ **Break**(M_k, M): **Break**(M_i, M) is minimal iff no other ϕ-model has a set of causal breaks that is included in the one of M_i. Similarly, **Intact**(M_i, M) is maximal among ϕ-models iff there is no ϕ-model M_k such that **Intact**(M_i, M) $\subset$ **Intact**(M_k, M). ϕ is an "atomic" proposition X = x. Thus:

ϕ-Minimal Model. Model M_i and **Break**(M_i, M) are ϕ-minimal relative to M iff

a. M_i is a ϕ-model;
b. for **Z**, the set of variables in G that are not descendants of ϕ, **Intact**(M_i, M) $\cap$ **Z** is maximal among ϕ-models;
c. **Break**(M_i, M) is minimal among ϕ-models.

Clauses (b) and (c) are a formalization of the idea that the counterfactual world should constitute a minimal departure from actuality as regards the noneffects of the counterfactual antecedent and that the changes introduced to make the antecedent true should be minimal: as minor and as late as possible, given the causal laws. Finally (cf. Hiddleston 2005, 643):

TCM (Theory of Counterfactuals in a Model).
($\phi > \psi$) is true in a model M and a context C iff ψ is true in every model M_i 1) that is ϕ-minimal relative to M and 2) for which **Break**(M_i, M) is relevant in C.

TCM roughly says that the counterfactual "Had ϕ been the case, ψ would have been the case" is true iff ψ follows, according to M's causal principles, from (a) ϕ itself, (b) events causally unrelated to ϕ and (c) actual causal relations ϕ does not prevent from obtaining. Context C determines what set of causal breaks is relevant to evaluate the counterfactual.

This theory of counterfactuals differs in important aspects from Lewis's and most other philosophers. Here I want to discuss two differences, both of which demonstrate that TCM is more useful for applications in the social sciences than that of the philosophers' tradition. First, counterfactual antecedents are implemented not by miracle—by breaking a law—but by *changing the value of a variable from actual to nonactual within the assumed system of causal principles.*

In the causal modeling literature that follows Lewis in this respect (e.g., Pearl 2000), a counterfactual antecedent is implemented by *changing* the laws. In particular, to evaluate whether a variable Y counterfactually depends on another variable X, all causal principles that have X as an effect are eliminated and replaced by the constant X = x. Doing so assumes that it is always possible to change causal principles one by one. In the semantics outlined here, causal principles remain the same.

As a consequence, second, counterfactuals will often backtrack. To see this, consider a slightly modified version of our example in which we now include a causal arrow from BFP to War, a binary variable describing whether or not the

Second World War happened. Let us suppose that BFP has indeed an influence on War so that:

$$(\text{BFP} = \text{dove}) \Rightarrow p(\text{War} = 1) = .75$$

$$(\text{BFP} = \text{hawk}) \Rightarrow p(\text{War} = 1) = .25.$$

Now, in order to evaluate the counterfactual "Had Britain taken a more confrontational stance in 1938, the likelihood of war would have been lower," we have to find a set of minimal causal breaks that make the antecedent true. There are four such sets: PM = {Churchill}, PM = {Eden}, PM = {Cooper}, and Cab = {hawk}. We cannot change BFP without changing either of its causes because the assumed causal principles do not allow such a change: In order for BFP to be different, either of its causes (or both) must be changed.

Therefore, the social scientists' criterion of historical consistency has a clear counterpart in this theory, but it is rendered more precise. Within the system of causal principles, a minimal change should make the antecedent, and "minimal" has a very precise meaning. Causal principles are not to be disturbed at all.

The remaining criteria can be accommodated within this theory. One can make a model more specific, for instance, by including more detail in the description of values for variables. Whether or not a given counterfactual is true is always determinate within a model. Cotenability is given by the causal principles. Moreover, depending on his or her orientation, a social scientist can demand that the causal principles follow from a theory or refuse to do so.

8.6. Four Problems for the Causal Theory of Counterfactuals

In this section I will discuss four problems that trouble the theory of counterfactuals that I have presented in the last section: the problem of circularity, the problem of backtracking, the problem of actual causation, and the problem of indeterminacy.

8.6.1. Circularity

Causal inference is not the only purpose of evaluating historical counterfactuals but it is an important one. The most obvious potential problem for the theory presented here is its circularity. If one needs a causal model in order to evaluate the counterfactual, doesn't one presuppose that the answer to the causal question is already known? This is certainly the case in above simplified example: BFP = hawk was *assumed* to have a direct causal influence on War = 1. It is therefore useless for causal inference.

However, the example was overly simplified. Adding slightly more structure shows that causal questions can have nontrivial answers, even if a lot of causal knowledge is presupposed. According to Khong, a confrontational British foreign policy would have influenced the occurrence of the Second World War through three separate routes (Khong 1996, 114–17). First, deterrence could have been successful, with Hitler backing down. Second, if unsuccessful, Hitler would have started a war but that would have triggered a coup d'état in turn, and the new German leaders would have sued for peace immediately. Third, Hitler would have jumped at the chance of war by invading Czechoslovakia, which would have started an earlier European war.

To model this scenario, we keep BFP for British foreign policy, distinguish two binary war variables, War_{38} for the war starting in 1938 and $War_{39\text{-}45}$ for the Second World War, and introduce two new binary variables, Det for whether or not deterrence was successful and CDE for whether or not a coup d'état was staged. The causal graph is illustrated in figure 8.2.

In this structure, it is not at all clear whether BFP makes a difference to $War_{39\text{-}45}$. This depends on the precise formulation of the causal principles, which, as mentioned above, may be indeterministic. Suppose the following principles are true:

$$(\text{BFP} = \text{hawk}) \Rightarrow p(\text{Det}=1) = .5;$$

$$(\text{BFP} = \text{dove}) \Rightarrow p(\text{Det} = 1) = 0;$$

$$(\text{Det} = 1) \Rightarrow p(War_{38} = 1) = p(War_{39\text{-}45} = 1) = 0;[\text{"Hitler backing down"}]$$

$$(\text{Det} = 0) \,\&\, (\text{BFP} = \text{hawk}) \Rightarrow p(War_{38} = 1) = 1;$$

$$(\text{Det} = 0) \,\&\, (\text{BFP}=\text{dove}) \Rightarrow p(War_{38} = 1) = 0;$$

$$(War_{38} = 1) \Rightarrow p(\text{CDE} = 1)=.7$$

$$(\text{CDE} = 1) \Rightarrow p(War_{39\text{-}45} = 1) = 0;[\text{"Coup"}]$$

$$(\text{CDE} = 0) \,\&\, (War_{38} = 1) \Rightarrow p(War_{39\text{-}45} = 1) = .7;[\text{"Hitler jumping at war"}]$$

$$(\text{Det} = 0) \,\&\, (War_{38} = 0) \Rightarrow p(War_{39\text{-}45} = 1) = .7.$$

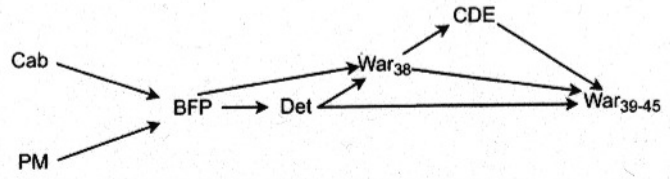

Figure 8.2 Many routes between BFP and $War_{39\text{-}45}$

Here I assume that the probability of a war starting in 1938 because of a confrontational British foreign policy but otherwise no different from World War II (i.e., $War_{38} = 1$ & $War_{39\text{-}45} = 1$) to be the same as the probability of the war as it happened (BFP = dove & $War_{39\text{-}45} = 1$) to represent the judgment that the "worst outcome [of confrontation] would have been 'no worse than the course of 1939.'" (Khong 1996, 117). For simplicity, other relations are deterministic except that confrontation may or may not have deterred Hitler, and that the coup d'état may or may not have happened given an earlier war.

In this model, with a confrontational course, the probability of war would have been .12, which is considerably lower than the .7 of appeasement. But this of course depends on the numbers. If we believe instead that an earlier war without a coup d'état would have made the Second World War *more* likely (.9), a coup d'état would have been overwhelmingly *un*likely (.1) and so was Hitler's backing down (.1), then a confrontational course would have increased the chances of war (from .7 to .729). The point is that these are not results that are directly built into the model or always obvious. The more complex the model, the more calculation one will need to reach conclusions.

8.6.2. Backtracking

I mentioned above that these semantics can force backtracking counterfactuals if causal principles of a certain kind are in place. Specifically, if the probability of a nonactual value of the variable that makes the counterfactual antecedent true, given its parents, is zero, only changing values for its parents can make the antecedent true (if anything). This may or may not lead to problems if causal inference is the purpose of evaluating the counterfactual.

The reason for Lewis and others to insist counterfactuals be nonbacktracking is that backtracking counterfactuals can lead to mistaken causal judgments. A stock philosopher's example will illustrate. Suppose you look at the barometer in your vestibule, see that its dials point to *very low*, and exclaim, "If only the hands were on *high*, there wouldn't have been a storm and I could have gone for my picnic!" In this context there is nothing wrong with this mode of counterfactual reasoning. What is required here is that the barometer provide evidence of the weather conditions, no matter for what reason. The reason that (a properly functioning) barometer is a reliable predictor of the weather conditions is that atmospheric pressure is a common cause of both the barometer reading and the weather conditions. Implicitly, the reasoning is therefore: If the dials were on high, atmospheric pressure would have been high, and therefore the weather would have been fair.

For exactly this reason backtracking counterfactuals can lead to counterexamples, if evaluating causal claims is the purpose. If, say, a child wanted to find out whether the barometer causes the storm (because she has observed that whenever the barometer said *low*, a storm would follow), it won't do to ask the counterfactual question "What would the weather conditions have been if the barometer reading had been different?" and to evaluate the counterfactual by backtracking. Counterfactual dependence is not a reliable indicator of causal connections when the counterfactual backtracks. To find out about the causation, the child would have to tinker with the barometer, change its reading in a way that breaks the causal connection with atmospheric pressure, observe whether the correlation persists, and only then make the causal judgment.

In the cases we focus on here, we are in a less fortuitous situation because we cannot tinker with the systems at hand. We resort to counterfactual speculation for precisely that reason. But if that is so, using backtracking counterfactuals as evidence for causal connections can lead to counterexamples. We can see that this is not a mere theoretical problem by adding yet another (plausible) causal pathway to our model: If Chamberlain had not been prime minister, Britain would have rearmed earlier.[14] The resulting causal graph is pictured in figure 8.3, where Arm is a binary variable indicating sufficient rearmament by 1938. In this structure, there can be counterfactual dependence of $War_{39\text{-}45}$ on BFP even though the likelihood of the former is unaffected by the latter (again, depending on the numbers of course).

The counterfactual thought experimenter who is interested in causal inference therefore seems to be in a dilemma: He either uses Lewis-style semantics but then winds up constructing historically inconsistent counterfactuals (in which, for example, Chamberlain confronted Germany after all), or he maintains historically consistent counterfactuals at the expense of incorrect causal judgments.

Either alternative is highly undesirable. Historical consistency isn't a mere academic prerequisite. In the tradition that goes back to Max Weber historical consistency is demanded in part for epistemic reasons: "Max Weber insisted that plausible counterfactuals should make as few historical changes as possible on the grounds that the more we disturb the values, goals, and contexts in which actors operate, the less predictable their behavior becomes" (Lebow 2010, 55). Thus, we seem to face the choice between an unreliable method for evaluating counterfactual dependence, which, if correct, reliably indicates causation, and a reliable method for evaluating counterfactual dependence, which is unreliable as an indicator of causation.

One way out of this dilemma consists in exploiting causal background knowledge—which is required to evaluate the counterfactual anyway. The definition TCM above mentions a context that determines what variables **Break**(M, M_i) are relevant, and so far I have not said more about what this context is and how it determines the

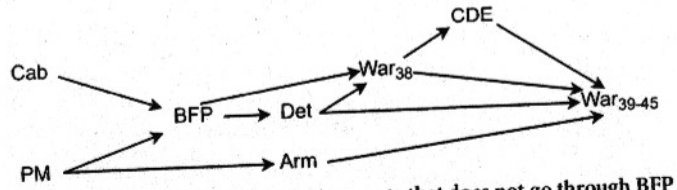

Figure 8.3 PM influences $War_{39\text{-}45}$ via a route that does not go through BFP

relevant causal breaks. Lewis mentions context dependence of counterfactuals in a passage that discusses backtracking (Lewis 1986, 34):

> (1) Counterfactuals are infected with vagueness, as everyone agrees. Different ways of (partly) resolving the vagueness are appropriate in different contexts. Remember the case of Caesar in Korea: had he been in command, would he have used the atom bomb? Or would he have used catapults? It is right to say either, though not to say both together. Each is true under a resolution of vagueness appropriate to some contexts. (2) We ordinarily resolve vagueness . . . in such a way that counterfactual dependence is asymmetric . . . Under this standard resolution, back-tracking arguments are mistaken . . . (3) Some special contexts favor a different resolution of vagueness, one under which the past depends counterfactually on the present and some back-tracking arguments are correct.

Lewis is mistaken to call the nonbacktracking resolution of vagueness ordinary or standard; it is just one resolution among others. In fact, there are good reasons to believe that ordinary language counterfactuals standardly backtrack. I gave the pressure-barometer-storm example above. Examples like this can be multiplied easily. They are particularly conspicuous in criminal investigations (which are, of course, not entirely unrelated to the historical counterfactuals we have been discussing). Here is a case that has the opposite structure of the barometer case. It is known that it was raining in the morning of the day of the crime. The detective sends forensic experts to look for tire marks near the crime scene. He reasons thusly: If there were tire marks, then if the suspect's car had been near crime scene, it would have to have stopped raining (as the car wouldn't have left marks if it hadn't stopped). Therefore, the car would have been at the crime scene in the afternoon.[15] The structure of this and the barometer case are as depicted in figure 8.4.

Here, then, are two contexts in which backtracking counterfactuals are permissible. Both are instances of evidential reasoning: We take the low (high) barometer reading as evidence for an oncoming storm (fair weather) just as we take, given the tire marks, the suspect's car being at the crime scene as evidence for the rain having stopped.

As is well known, causal reasoning differs from evidential reasoning. No one would take the evidential connection between the barometer reading and the storm or between the suspect's car being at the crime scene and the rain having stopped, given the tire marks, as indicating a causal connection. Causal analysis therefore differs from these contexts.

I propose to amend TCM as follows. In the context of causal analysis, **Break**(M, M_i) may only contain variables that are connected to the putative effect variable, if

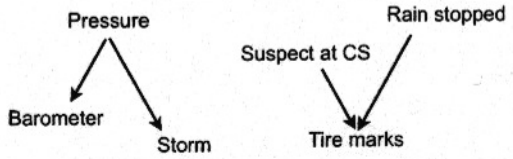

Figure 8.4 Backtracking counterfactuals

at all, only through directed paths that include the putative cause variable.[16] By a directed path I mean an ordered sequence of variables A, D_1, D_2, . . . , D_n, C such that an arrow goes from A to D_1, from D_1 to D_2 and so on to C. In our example, there is a directed path from PM to $War_{39\text{-}45}$ via BFP and Det, one via BFP, Det and War_{38} and so on, and, importantly, one via Arm. BFP is the putative cause variable and $War_{39\text{-}45}$ the putative effect variable. Therefore, there is one path, PM → Arm → $War_{39\text{-}45}$, that does not include the putative cause variable. In the context of causal analysis, this is an inadmissible counterfactual.

However, there is another variable, Cab, that is connected to $War_{39\text{-}45}$ only through directed paths that contain BFP. Khong includes this alternative way to make the antecedent true because he is not so sure about whether it is indeed the case that someone other than Chamberlain could have been prime minister; that is, he is uncertain as to whether nonactual values of the variable PM have a positive probability.[17] Khong wants to be on the safe side of historical consistency. I have argued that the existence of this additional cause is fortuitous from a methodological point of view: Without it, counterfactual dependence of the Second World War on British foreign policy would not be a good indicator of a causal dependence. It is evident, however, that nothing guarantees that such a variable will always be available.

8.6.3. Actual Cause

As mentioned in section 8.2, the counterfactual theory of causation is plagued by counterexamples that involve cases of redundant causation in which two or more causes compete in their bringing about an effect. Billy and Suzy throw rocks at a bottle. As it happens, Suzy's rock hits the bottle first, shattering it into 1,000 pieces. But there is no counterfactual dependence: Had Suzy not thrown her rock (or had it not hit the bottle first), Billy's rock would have shattered the bottle anyway. Saying that the shattering caused by Suzy's rock is a different event from the shattering that would have been caused by Billy's rock may help in this case (though I doubt it) but describing an event in more detail as a strategy does not work in general (see section 8.4.1).[18]

In Reiss 2009 I argued that social scientists, by and large, are not interested in causes that do not make a difference. Suppose that appeasement was indeed an actual cause of the Second World War: Hitler perceived the UK leaders as weak and took their weakness as reason to occupy the remainder of Czechoslovakia after annexing the Sudetenland and eventually to attack Poland. But suppose too that there was a second psychological mechanism waiting in the offing that would have caused him to order the occupation of Czechoslovakia *despite* Britain's threat ("now more than ever!"). In this hypothetical scenario, appeasement was the actual cause of war but it wasn't a cause that made a difference.

One purpose of evaluating counterfactuals is to provide insight into policy choices (Reiss and Cartwright 2004). To the extent that policy analysis is the goal of the investigation, finding factors that do not make a difference to the outcome of interest is not very useful. Thus, while the counterfactual approach can and will

lead to misjudgments about *actual* causes, this problem is benign from a policy point of view. Historians and social scientists look for *causes that make a difference*, not for actual causes.

8.6.4. Indeterminacy

The fourth and final problem is the most serious of the four. In fact, as far as I can see there is no solution to it within the present framework. Social scientists and historians often use qualifiers such as "likely," "probably," "the chance that," and so on. In TCM, such uncertainty can be modeled by assigning probabilities to outcomes that lie strictly between zero and one. But if even one link on the route from antecedent to consequent is indeterministic, the probability of the consequent will be strictly between zero and one. That, in turn, makes all "would" counterfactuals false.

Consider the example discussed in section 8.6.1. Here the probability of the Second World War is .12, given BFP = hawk. However, in this scenario the counterfactual "Had Britain confronted Germany, the Second World War would have been avoided" is obviously false. By contrast, the counterfactual "Had Britain confronted Germany in 1938, the Second World War *might* have been avoided" is true. But this is not very informative, especially given the probability of war given appeasement was also below unity (that is, the counterfactual "Had Britain tried to appease Germany in 1938, the Second World War might have been avoided" is also true).

We could say that the model sustains the counterfactual "Had Britain confronted Germany in 1938, the Second World War would *probably* have been avoided." But then we would encounter a threshold problem: How small does the probability of an outcome have to be in order for an outcome being "probable not to have happened" (and vice versa for high probabilities)? A possible threshold is .5: Outcomes with a probability below .5 are probable not to have happened, above .5 they are probable to have happened. What about .5 itself? Should we say that the event is both probable to have happened *and* probable not to have happened? This is obviously awkward, but arbitrarily deciding either way is just as awkward. Similarly awkward choices have to be made for any threshold level.

Moreover, even if one could settle the threshold question, the "probably" counterfactual would still not be very informative. Suppose the counterfactual "Had Britain confronted Germany in 1938, the Second World War would *probably* have been avoided" was evaluated by TCM and is indeed true. To learn this claim is only useful to the extent that the alternative policy made the Second World War very likely. We are interested in whether or not Britain's foreign policy *made a difference* to the likelihood of war, not in the likelihood of war per se.

We are now moving close to a probabilistic theory of causation. We could for example define:

> PTM (Probabilistic Theory of Causation in a Model)
> $X = x$ rather than $X = x'$ causes $Y = y$ iff $P(Y = y \mid X = x) > P(Y = y \mid X = x')$ in an $(X = x)$-minimal model.

It is beyond the scope of this chapter to discuss the probabilistic theory. Suffice it to say that variables that *lower* the probability of an outcome can be its causes. This can happen whenever an outcome occurs *despite* the existence of a preventer. Suppose that Britain did indeed confront Germany in 1938. Hitler could not be deterred and the war breaks out early. A coup d'état is staged but it fails. The Second World War happens almost as it did except for an earlier beginning. In this scenario confrontation lowers the probability of $War_{39\text{-}45}$ (using the figure from section 8.6.1) but it is among the causes of the war.

8.7. Implications: Counterfactuals and Purpose

Whether or not counterfactual speculation is a worthwhile activity depends on the purpose and the specific facts of the case being speculated about. This chapter has concentrated on one salient purpose, causal inference. I have argued that there are four major problems in the way of using the counterfactual account for causal inference. Of the four, I argued that the fourth—the problem of indeterminacy—is likely to be the most damaging: To the extent that some of the causal principles that connect counterfactual antecedent and consequent are genuinely indeterministic, the counterfactual will be of the "might have been" and not the "would have been" kind. I want to finish with some observations regarding these might have been counterfactuals.

Jon Elster made a very perceptive remark in his *Logic and Society* (Elster 1978, 184–5):

> One crucial aspect is that the theory T emerges as something more than just an instrument that permits us to conclude from the hypothetical antecedent to the hypothetical consequent: it also serves as a filter for the acceptance or the rejection of the antecedent itself. Thus for a successful counterfactual analysis a delicate balance must be struck: the theory must be weak enough to admit the counterfactual assumption, and also strong enough to permit a clear-cut conclusion.

Here I have focused on parts of the social science with few accepted theories such as history and international relations but a similar observation holds: The causal principles describing a situation of interest must be weak enough—that is, contain genuinely indeterministic relations so that the counterfactual antecedent can be implemented. If there was no hawk in British politics in 1938, Britain could never have confronted Germany. At the same time, the principles must be strong enough—that is, contain enough deterministic relations so that the consequent follows from the antecedent together with the principles. Using the semantics of section 8.5, we can thus make Elster's observation more precise: What is required is enough indeterministic causal relations so that the antecedent can be implemented and enough deterministic relations so that the consequent (or its negation) follows.

Evidently, this is a tall order: Why would deterministic and indeterministic causal principles be distributed in just this way? Wouldn't it seem likely that to the extent we are willing to believe that the antecedent event was contingent, we are also willing to believe that the outcome remained contingent given the antecedent event? Contrapositively, wouldn't it seem likely that to the extent we are willing to believe that a consequent had to follow given the antecedent, we also believe that the antecedent was necessary to begin with?

Despite this in my view very serious problem of the counterfactual account, counterfactual speculation nevertheless has some virtue. First, if it is conducted in accordance with strict rules such as those described in section 8.5, a lot can be learned in the process of building a causal model. The causal principles necessary to implement the antecedent and evaluate the consequent cannot be read off standard historical accounts of a given situation. Judgments concerning causal relations among events of interest and their probabilities given the causal parents must be supplied with evidence. One of the advantages of the formal apparatus introduced here is that the apparatus provides definite guidelines regarding the required information for evaluating a counterfactual. Another advantage is that once a model has been built disagreements can be made explicit and specific. One might disagree with the counterfactual "Had Britain confronted Germany in 1938, the Second World War would have been no worse and most likely would have been less detrimental than it was" because one denies that Britain could have confronted Germany, or because one denies that the three routes through which Britain's policy affected the war existed. If one denies that Britain could have confronted Germany, one would have to supply an argument to the effect that a confrontational politician such as Churchill could not have been prime minister at the time. One would have to supply an alternative model in which Churchill's preferred course of action was inevitable, and one would have to establish that this is the better model of the situation.

Second, counterfactuals are useful for purposes other than causal inference. One such purpose is the reduction of cognitive bias. Social scientists tend to regard the future as open and contingent but the past as inevitable: "Work on hindsight bias shows that as soon as observers learn the outcome of an historical process they begin to reorganize their understanding of the causal forces at work so that the outcome appears more retrospectively foreseeable than it was prospectively" (Tetlock and Parker 2006, 25). Having to build an explicit causal model for a situation will force commentators to make explicit judgments about the probabilities of events given outcomes and it is likely that even someone who will initially regard the First or Second World War as inevitable will not judge all causal principles to be deterministic. By that means, perceptions of outcomes as being inevitable will be reduced and hindsight bias decreased.

A final argument in favor of counterfactuals even in the context of establishing causation is that there are no alternatives that are unequivocally superior. The main alternative to the counterfactual account is process tracing. But process tracing is itself not without problems. One issue is conceptual: Process tracing establishes whether a factor is causally connected to an outcome but this may not at all be what we want to know. As argued above, social scientists tend to be interested in causes that actually made a difference, and a factor's being causally connected to the outcome does not entail that it made a difference to it. For instance, a factor might be causally connected to an outcome through various routes, some of which promoted the outcome, some prevented it, and knowing just about the links does not tell us whether the overall contribution was positive, negative, or nil. The second issue is epistemic. Process tracing works best at the individual level and requires a great deal of knowledge about actors' motivations, goals, and deliberations, which cannot always be had. Even if, for instance, there are records of members of staff reporting that some decision was made for such-and-such a reason, the politician in question may just have said so because that was the politically acceptable, but not necessarily actual, reason. For all its difficulties, counterfactual speculation may sometimes be the only way to make causal inferences about singular events.

ACKNOWLEDGMENTS

I would like to thank the participants of the the Philosophy and Methodology of the Social Sciences held at the University of Alabama Birmingham in April 2010 for valuable comments, and especially Harold Kincaid for putting the conference together and his comments on this paper. I would also like to thank audiences at the Philosophy of Social Science Roundtable (Paris, March 2011), at a workshop on causation held in Dijon in June 2010, and students at Erasmus University Rotterdam for their comments and suggestions. Financial support from the Spanish government (research projects FFI2008-01580 and CONSOLIDER INGENIO CSD2009-0056) is gratefully acknowledged.

NOTES

1. Much more detail needs to be added to make the account plausible even on the surface. These details do not matter, however, to the present discussion.

2. It is vacuously true iff there are no $\neg C$-possible worlds.

3. Most counterexamples involve statistical mechanical or other indeterministic scenarios, see Jackson 1977; Elga 2000; Schaffer 2004; Hawthorne 2005; Noordhof 2005.

4. In fact, it reads "Had President Kennedy issued a timely warning in the spring of 1962, Khrushchev *might* not have sent missiles to Cuba" (Lebow and Stein 1996, 124; emphasis added). I will discuss the difference between "would" and "might" counterfactuals below, and will not complicate the discussion unnecessarily here.

5. Elster 1978 criticizes Robert Fogel and other so-called new economic historians partly on similar grounds. Fogel 1964, for instance, evaluates the social savings the

United States made relative to a no-railway economy without asking whether such an economy could have existed in the late nineteenth century. See the more detailed discussion below.

6. Cf. Lebow 2010, 54; Tetlock and Belkin 1998, 19.

7. In some recent work on the counterfactual account (e.g., Schaffer 2005; Northcott 2008; Reiss 2011) causation is made explicitly contrastive; that is, causal relations are three or four place and of the form "*C* rather than *C** causes *E*" or ". . . *E* rather than *E**". Thus, Susan's *stealing* the bike (rather than buying it) caused her to be arrested; but Susan's stealing *the bike* (rather than the skis) did *not* cause her to be arrested. The problem of clarity reappears when contrast events are described. Perhaps had Susan bought the bike using counterfeit money or a stolen credit card, she would have been arrested anyway.

8. This problem is recognized among philosophers, who discuss it in connection with the nature of eventhood. They call an event "fragile" to the extent that small differences in time, place, or manner of occurrence make for a numerically different event and realize that events must be understood as having the appropriate degree of fragility in order to avoid counterexamples. Collins, Hall, and Paul (2004, 44) therefore rightly argue that "the theory of events thus counts as a subtheory of a complete theory of causation." There is nevertheless a difference between the philosophers' and the social scientists' treatment of this question. Whereas philosophers tend to presume that there is one correct theory of eventhood that covers all cases, social scientists merely demand that the details of cause and effect be specific enough to get the particular case right.

9. Cf. Tetlock and Belkin 1996, 21; Lebow 2010, 55.

10. Cf. Tetlock and Belkin 1996, 23; Lebow 2010, 55; Weber 1949 [1905]; Hawthorn 1991, 158.

11. Cf. Tetlock and Belkin 1996, 25–30; Lebow 2010, 56. Also cf. for instance: "there are no 'law-like' and few well-established statistical generalizations in the field of international relations" (Lebow and Stein 1996, 127; see also Lebow 2010, 56).

12. By that of course I mean the structural realists in international relations, not those in the philosophy of science.

13. The influence of the cabinet on foreign policy I take from the following passage: "My argument does not depend on one of them being prime minister in 1938. Had two or more of the Churchill-Eden-Cooper trio been members of the Chamberlain cabinet in September 1938, the chances of Britain's confronting Hitler would have greatly increased" (Khong 1996, 113–14).

14. "Perceived military weakness in 1938 definitely stayed Chamberlain's hand, but Chamberlain's handiwork—both as chancellor of the exchequer and as prime minister—was also responsible for Britain's military underpreparedness" (Khong 1996, 101).

15. The variable Tire Marks is what in the causal modeling literature is called a "collider." It is well known that conditioning on a collider creates dependencies among its parents.

16. Such a variable is akin to an instrumental variable in econometrics. See, for instance, Reiss 2008, chapter 7.

17. "For those dubious about whether any one of the trio could have been prime minister, in 1938, this respecification of just having two or more of them as cabinet ministers in 1938 might be closer than the original specification of one of them as prime minister to the counterfactual world in which a confrontational Britain challenges Germany" (Khong 1996, 114).

18. This problem is recognized in the law; see, for instance, Fumerton and Kress 2001.

REFERENCES

Ball, L., and G. Mankiw. 2002. "The NAIRU in Theory and Practice." NBER Working Paper 8940. Cambridge, MA: National Bureau of Economic Research.

Carr, E. 1986 [1961]. *What Is History?* 2d ed. Houndmills: Macmillan.

Cartwright, N. 2007. "Counterfactuals in Economics: A Commentary." In *Hunting Causes and Using Them*, 236–61. Cambridge: Cambridge University Press.

Collins, J., N. Hall, and L. Paul. 2004. *Causation and Counterfactuals*. Cambridge, MA; The MIT Press.

Cummings, R. 2006. "'What if': The Counterfactual in Program Evaluation." *Evaluation Journal of Australasia* 6 (2): 6–15.

Dawid, P. 2000. "Causal Inference Without Counterfactuals." *Journal of the American Statistical Association* 95 (45): 407–27.

Elga, A. 2000. "Statistical Mechanics and the Asymmetry of Counterfactual Dependence." *Philosophy of Science* 68 (Supplement): 313–24.

Elster, J. 1978. *Logic and Society: Contradictions and Possible Worlds*. Chichester: John Wiley.

Ferguson, N. 1997. *Virtual History: Alternatives and Counterfactuals*. New York: Basic Books.

Fine, K. 1975. "Critical Notice: David Lewis' Counterfactuals." *Mind* 84 (1): 451–58.

Fischer, D. H. 1970. *Historians' Fallacies: Toward a Logic of Historical Thought*. New York: Harper and Row.

Fogel, R. 1964. *Railroads and American Economic Growth*. Baltimore: Johns Hopkins University Press.

Fumerton, R., and K. Kress. 2001. "Causation and the Law: Preemption, Lawful Sufficiency, and Causal Sufficiency." *Law and Contemporary Problems* 64 (4): 83–105.

Goldthorpe, J. 2001. "Causation, Statistics, and Sociology." *European Sociological Review* 17 (1): 1–20.

Gopnik, A. 2009. *The Philosophical Baby*. London: Bodley Head.

Gopnik, A., C. Glymour, D. Sobel, L. Schulz, and T. Kushnir. 2004. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets." *Psychological Review* 111 (1): 3–32.

Hawthorn, G. 1991. *Plausible Worlds: Possibility and Understanding in History and the Social Sciences*. Cambridge: Cambridge University Press.

Hawthorne, J. 2005. "Chance and Counterfactuals." *Philosophy and Phenomenological Research* 70 (2): 396–405.

Heckman, J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35 (1): 1–97.

Hellekson, K. 2001. "The Alternate History: Refiguring Historical Time." Kent, OH: Kent State University Press.

Hiddleston, E. 2005. "A Causal Theory of Counterfactuals." *Noûs* 39 (4): 632–57.

Hitchcock, C., and J. Knobe. 2009. "Cause and Norm." *Journal of Philosophy* 106 (11): 587–612.

Holland, P. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Hume, D. 1777 [1902]. *Enquiry Concerning Human Understanding*. L. A. Selby-Bigge, ed. Oxford: Clarendon Press.

Jackson, F. 1977. "A Causal Theory of Counterfactuals." *Australasian Journal of Philosophy* 55 (1): 3–21.

Khong, Y. F. 1996. "Confronting Hitler and Its Consequences." In *Counterfactual Thought Experiments in World Politics*, P. Tetlock and A. Belkin, eds., 95–118. Princeton, NJ: Princeton University Press.

Lebow, R. N. 2010. *Forbidden Fruit: Counterfactuals and International Relations*. Princeton, NJ: Princeton University Press.
Lebow, R. N., and J. G. Stein. 1996. "Back to the Past: Counterfactuals and the Cuban Missile Crisis." In *Counterfactual Thought Experiments in World Politics*, P. Tetlock and A. Belkin, eds., 119–48. Princeton, NJ: Princeton University Press.
Lewis, D. 1986. "Counterfactual Dependence and Time's Arrow." In *Philosophical Papers*, vol. II, 32–51. Oxford: Oxford University Press.
Lewis, D. 2000. "Causation As Influence." *Journal of Philosophy* 97 (4): 182–97.
Markle, G., and F. McCrea. 2008. *What If Medicine Disappeared?* Albany: State University of New York Press.
Maudlin, T. 2007. *The Metaphysics Within Physics*. Oxford: Oxford University Press.
McDermott, M. 1995. "Redundant Causation." *British Journal for the Philosophy of Science* 46 (4): 523–44.
Mohr, L. 1995. *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: Sage.
Morgan, S., and C. Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.
Morton, R., and K. Williams. 2010. *Experimental Political Science and the Study of Causality*. Cambridge: Cambridge University Press.
Noordhof, P. 2005. "Morgenbesser's Coin, Counterfactuals and Independence." *Analysis* 65 (3): 261–63.
Northcott, R. 2008. "Causation and Contrast Classes." *Philosophical Studies* 139 (1): 111–23.
Pearl, J. 2000. *Causation: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
Reiss, J. 2007. "Do We Need Mechanisms in the Social Sciences?" *Philosophy of the Social Sciences* 37 (2): 163–84.
Reiss, J. 2008. *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge.
Reiss, J. 2009. "Counterfactuals, Thought Experiments and Singular Causal Analysis in History." *Philosophy of Science* 76 (5): 712–23.
Reiss, J. 2012. "Causation Isn't Contrastive, It's Contextual." *Philosophy Compass*. Forthcoming.
Reiss, J., and N. Cartwright. 2004. "Uncertainty in Econometrics: Evaluating Policy Counterfactuals." In *Economic Policy Under Uncertainty: The Role of Truth and Accountability in Policy Advice*, P. Mooslechner, H. Schuberth, and M. Schürz, eds., 204–32. Cheltenham: Edward Elgar.
Rubin, D. 1974. "Estimating the Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.
Rubin, D. 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2 (1): 1–26.
Schaffer, J. 2004. "Counterfactuals, Causal Independence and Conceptual Circularity." *Analysis* 64 (4): 299–309.
Schaffer, J. 2005. "Contrastive Causation." *Philosophical Review* 114 (3): 327–58.
Sloman, S. 2005. *Causal Models: How People Think About the World and Its Alternatives*. Oxford: Oxford University Press.
Tetlock, P., and A. Belkin, eds. 1996. *Counterfactual Thought Experiments in World Politics: Logical, Methodological and Psychological Perspectives*. Princeton, NJ: Princeton University Press.
Tetlock, P., and G. Parker. 2006. "Counterfactual Thought Experiments: Why We Can't Live Without Them & How We Must Learn to Live With Them." In *Unmaking the West: "What-If" Scenarios that Rewrite World History*, P. Tetlock, R. N. Lebow, and G. Parker, eds., 14–44. Ann Arbor: University of Michigan Press.
Tetlock, P., R. N. Lebow, and G. Parker. 2006. *Unmaking the West: "What-If" Scenarios that Rewrite World History*. Ann Arbor: University of Michigan Press.
Thompson, E. 1978. "The Poverty of Theory: or an Orrery of Errors." In *The Poverty of Theory and Other Essays*, 193–398. London: Merlin.
Weber, M. 1949 [1905]. *Objective Possibility and Adequate Causation in Historical Explanation. The Methodology of the Social Sciences*, M. Weber, E. Shils, and H. Finch, eds., 164–88. New York: The Free Press.