justify our political biases and predilections? There appear to be large classes of questions in the study of global conflict and cooperation for which experimental control is out of the question and statistical control is of limited usefulness (assuming we can find a reasonable set of comparison cases and can reliably operationalize the theoretical constructs). These questions are too important to ignore, but apparently too difficult to answer in a fashion that commands transideological consensus.

Too often, the response to the dilemma is to embrace extreme solutions (Strassfeld 1992): either to reject categorically all counterfactual arguments as fanciful suppositions, mere conjecture, and frivolous figments (counterfactual dread) or to assume confidently that we know exactly what would have happened if we had gone down another path, sometimes going so far as to project several steps deep into hypothetical causal sequences (counterfactual bravado). The former response leads to futile efforts to exorcize counterfactuals from historical inquiry (Fisher 1970); the latter response leads at best to error (we ignore the compounding of probabilities at our peril) and at worst to the full-scale politicization of counterfactual argument (as advocates claim carte blanche to write hypothetical histories that advance their favorite causes). This book tries to articulate a principled compromise between these extremes. On the one hand, we acknowledge that thought experiments inevitably play key roles in the causal arguments of any historical discipline. On the other hand, we acknowledge that thought experiments are often suffused with error and bias. But, that said, we do not conclude that things are hopeless—that it is impossible to draw causal lessons from history. Rather, we conclude that disciplined use of counterfactuals—grounded in explicit standards of evidence and proof—can be enlightening in specific historical, theoretical, and policy settings. And that, we suspect, is the most important lesson of this book.

2

Causes and Counterfactuals in Social Science

EXPLORING AN ANALOGY BETWEEN CELLULAR AUTOMATA AND HISTORICAL PROCESSES

JAMES D. FEARON

FOR A VARIETY of purposes, social scientists and historians take the discovery of causes of events in the human world as a goal—perhaps the principal goal—of their work.[1] Some research communities are shy of the word "causes," preferring words like "influences," "determinants," "sources," "origins," "roots," "correlates," "factors that shape or give rise to," and so on. But these are all forms of language that is basically causal.[2]

When trying to argue or assess whether some factor A caused event *B*, social scientists frequently use counterfactuals.[3] That is, they either ask whether or claim that "if *A* had not occurred, *B* would not have occurred." Most often, such claims are little more than unelaborated rhetorical devices—throwaway lines—deployed as part of a larger rhetorical strategy to convince the reader that *A* caused *B*. Less frequently, researchers actually develop and explore the counterfactual scenario as a means of testing the causal hypothesis.

Whether counterfactual argument should be considered a valid method of testing causal hypotheses is not clear. Considerable skepticism has been expressed over the years, focusing on the objection that it is difficult or impossible to know with any certainty what would have happened if some proposed cause had been absent in a particular historical case. This is a strong objection. Who can say with any assurance what would have happened if

[1] I wish to thank Mark Hansen, David Laitin, participants of the conference "Counterfactual Thought Experiments in World Politics," and especially Aaron Belkin, David Collier, Philip Tetlock, Mark Turner, and Peter Woodruff for valuable comments.

[2] I am aware that there is significant debate among philosophers about whether valid explanations are all "causal" and have the same basic form, or fundamentally differ, for example, from causal explanations of physical events to intentional explanations of actions. When I say that all social scientists seek to discover causes, I do not mean that they all think about causes the same way; I would like to include intentionalist explanations, however they are precisely characterized. On the debate see Wright (1971) and Davidson (1980).

[3] See Tetlock and Belkin (Chapter 1) and Fearon (1991).

Neville Chamberlain had not pursued a policy of appeasement, or if nuclear weapons had not been invented, or if the Reagan administration had not engaged in such a large defense buildup?

Nonetheless, there are also reasons to believe that social scientists, who generally cannot conduct true experiments, may have no choice but to rely on counterfactual assertions in one way or another. If some event *A* is argued to have been the cause of a particular historical event *B*, there seems to be no alternative but to imply that a counterfactual claim is true—if *A* had not occurred, the event *B* would not have occurred. And it can be shown that causal claims evaluated with regression and related methods applied to nonexperimental data must assume the truth of a counterfactual proposition concerning other causes of the phenomenon in question (Fearon 1991, 174–75).[4]

In this chapter I focus on two problems that bear on the questions of whether and how counterfactuals should be used by social scientists. The first is the objection noted above: How can we know with any confidence what would have happened if the hypothesized causal factor had been absent? I argue that for most social science problems we simply cannot know and, moreover, we cannot know in principle. Further, the difficulties involved in peering into possible worlds put fairly strong constraints on how much solid empirical confirmation we can get from any conceivable method of counterfactual argument.

Nonetheless, although it is frequently impossible to say with much precision what would have happened if A had been different, it is often easy and plausible to argue the negative case that whatever would have happened, it would *not* have been *B*, and therefore that *A* was a "cause" of *B*. This raises the second problem. For *too many* factors *A* it can be plausibly argued that but for *A*, *B* would not have occurred. How do we select among all the possibilities? Why is it not totally arbitrary to select one factor and argue for its causal status on counterfactual grounds if similar arguments can be advanced for myriad other factors and events? Are there criteria by which some counterfactual antecedents should be judged legitimate while others should not be considered because they are illegitimate to vary counterfactually as potential causes? For example, should we say that the length of Cleopatra's nose was a cause of World War I, and Napoleon's lack of a Stealth bomber a cause of his defeat at Waterloo? Or that it is illegitimate to consider the 1930s without a British policy of appeasement because Chamberlain could not have pursued any other policy given the constraints he faced?

This second problem, I will suggest, turns on ambiguity or confusion about the meaning of the word "cause" as we understand it when discussing historical phenomena. I offer two arguments. First, when we say that "*A* caused *B*" we seem to mean not just that if *A* had not occurred, *B* would not have occurred. Rather, we mean that if *A* had not occurred, *B* would not have occurred and the world would otherwise be similar to the world that did occur. This takes care of the Cleopatra's nose example and other factors that might be argued to be causes on "butterfly effect" grounds.

My second argument is that what we understand by "cause" differs in different explanatory environments and problems.[5] In particular, I argue that what we accept as a cause differs according to whether we are trying to give causes of a particular historical event, such as World War I, or of a class of events, such as wars in general. In the case of particular events, we often seek what I will call *conceivable causes*, factors that could actually have been different, according to the best of our knowledge about how the social and physical worlds work. On the other hand, when we argue that some factor causes some event (such as war) across cases, we do not typically require that in each case it be actually or "objectively possible" that the factor not occur. I call such factors *miracle causes*.[6] For example, one might maintain that imbalances of power cause war across cases and use the late 1930s in Europe as a supporting case, even if historians claim that the British and French could not conceivably have rearmed faster than they did. Thus, giving the causes of a singular event and of a class of events may be different sorts of explanatory exercises, and what can be "legitimately" accepted as a counterfactual antecedent may differ according to the exercise. One implication is that a researcher's purpose of inquiry may reasonably determine what factors should or should not be varied counterfactually, and thus what the causes of the phenomenon are!

In developing these arguments, I have found it helpful to use a model known as a cellular automaton as an analogy for the sort of historical phenomena about which social scientists make causal arguments. Independent of the arguments sketched above, this analogy is valuable for thinking about counterfactuals and historical processes.

Cellular Automata

The following is an example of a two-dimensional cellular automaton: Imagine a computer screen divided by a grid into a large number of cells, for example, one hundred cells on each side. In each of a successive number of

[4] For arguments that social scientists and historians cannot avoid relying on counterfactuals see Fogel (1964), Fearon (1991), and Tetlock and Belkin (Chapter 1).

[5] This general observation is made by Hart and Honoré (1959, 17), who find the first treatment of it in Mill (1900).

[6] The idea of a "miracle" explaining how things could have been different comes from Lewis (1973, 76).

periods $t = 0, 1, 2, 3, \ldots$, every cell on the screen will take on one color from a set of colors—in our simple version, there are two possible colors, green or yellow. Next, there is a rule that determines what color a cell will be in period t as a function of its own color and the colors of its immediate neighbors in period t-1. For instance, a rule might say "green if two or three neighbors were yellow, yellow otherwise."

If the rule is deterministic—that is, the rule determines cell colors with certainty rather than with some probability—then given any initial ($t = 0$) distribution of cell colors the system will evolve along a deterministic path. Even for very simple rules, however, it may be impossible to write down an equation that will give the color of a given cell in a given period t as a function of the initial pattern of colors. It should be stressed that this is not due to any random element in the automaton. Rather, there simply does not exist a formula or any other simplified model of the system that can project the system's behavior. As Stephen Wolfram (1984, 32) puts it, for some transition rules the behavior of a cellular automaton is "essentially unpredictable, even given complete information about the initial state: the behaviour of the system may essentially be found only by explicitly running it."

Early computer experiments with cellular automata showed that even quite simple deterministic rules could generate highly elaborate images that appear over time to move, grow, shrink, envelop, explode, spiral, "eat," and contort all over the screen. Tiny differences in the initial pattern might have enormous implications later on. For example, for many rules, changing a single cell from yellow to green in period $t = 0$ would mean that one hundred generations later the pattern would be unrecognizably different. For many rules there is no long-run equilibrium pattern that the system gradually evolves to regardless of the initial state. Systems may constantly change and evolve, like a pattern of sunlight through the leaves of a tree on a windy day.[7]

We owe the idea of a cellular automaton to the mathematicians John Von Neumann and Stanislaw Ulam.[8] While thinking about computers, Von Neumann apparently had some ideas about self-reproduction in biological systems, which he tried to express via cellular automata. The best-known example of an interesting rule for a two-dimensional cellular automaton is

[7] Of course, in a finite cellular automaton there is only a finite number of possible patterns for the screen, so all initial patterns either converge to some equilibrium or end up cycling through a finite set of patterns. For a large automaton, however, the number of possible patterns is enormous, so a "cycle" can easily take longer to complete than anyone would have time to observe.

[8] For Von Neumann's essays, see Von Neumann (1966). For a more recent overview see Wolfram (1983). This paper, along with other technical works on cellular automata, are collected in Wolfram (1986). Thus far I have only been able to find a technical literature on automata and a small, computer-hobbyist, mathematical-games literature. For the latter see Gardner (1970) and, more recently, Sigmund (1993). In both literatures authors almost invariably note that the behavior of automata seems to mimic natural, historical processes, but I have yet to see this analogy developed in more than a passing comment.

"The Game of Life," created by John Conway. Using the above example, the transition rule for the Game of Life is: (1) a cell is yellow in period t if less than two or more than three of its neighbors were green in the last period; (2) if exactly two neighbors were green in the last period, the cell stays the same color it was; and (3) if exactly three neighbors were green, a cell turns green or stays green. The interpretation Conway suggests is that green represents a living organism, while yellow represents a dead one or an empty space. The logic behind the rule is that an organism needs a certain number of living neighbors to survive or be born, but too many yields overcrowding, resource depletion, and death. The Game of Life can then be thought of as a model of the evolution of a bacteria population in a petri dish, for instance.

The Game of Life yields dynamic behavior like that described above. There may be no long-run equilibrium state and the patterns that evolve are very sensitive to initial conditions. Self-reproducing "structures" (i.e., patterns within the whole pattern) are typically generated and may endure for many periods, until they encounter other structures that may absorb or disintegrate them. Note that all this variety, complexity, and chaos can follow from a very simple, deterministic transition rule.

Cellular automata of this sort provide an appealing and fruitful analogy to the historical processes studied by historians and social scientists.

The Analogy

Imagine a large and fairly complex cellular automaton, in which the cells can assume many different colors and for which the transition rules are stochastic rather than deterministic. That is, part of a rule might specify something like the following: if exactly two neighbors were green in the last period, then the cell will be green with probability .3 and yellow with probability .7 in the current period. With a stochastic rule, the path of the patterns that evolve from an initial state will obviously no longer be deterministic. Instead, there will be a hopelessly complex implied probability distribution on possible patterns (which number $2^{10,000}$ in the two-color, 100-by-100 example used above). Now it is not even clear how much one learns from observing a single simulation. At each period t, there will be a range of actually possible successor patterns in period $t+1$, with chance deciding which occurs. And as in the deterministic example, if a single cell just happens to flip green rather than yellow in one period, this can have enormous consequences one hundred or even ten periods later.[9]

[9] Why should an automaton that is analogous to the human world be pictured as stochastic rather than deterministic? We might think that (1) the world is "truly stochastic" at some micro, atomic level; (2) mechanisms of human choice sometimes involve randomization, or "random-

An event in the social world is analogous to the appearance, disappearance, or change of some pattern within this automaton. For example, a war might be analogous to the appearance of a ring of red cells against a blue background, while the disintegration of a government might be analogous to the disappearance or fragmentation of a cluster of black cells. Each "event" might have its own historically unique aspects. The colors of nearby cells might be novel, or the ring might be shaped, shaded, and growing in ways that were slightly different from any previous ring that had been observed. But nonetheless it might still be sensible to speak of recurrent patterns that are identifiable as such (e.g., red rings). In other words, because every "case" might look different in particular respects, it might be difficult to code some of them, but nonetheless we could recognize categories.

A transition rule is analogous to a set of causal mechanisms that explain social, political, or economic interactions. For example, perhaps the most commonly employed mechanism in historical explanation is: People choose actions that make sense ("are optimal") in light of their beliefs and objectives. The various psychological biases in decision making discovered by cognitive psychologists provide another class of examples. A causal mechanism might also be less microlevel—it could be constructed or based on such microlevel components. For instance, the proposition that states will fight wars when both sides are overly optimistic about their chances of winning might be seen as a transition rule that is built up from more primitive rationality or psychological-bias mechanisms.

What follows if historical processes are "like" a stochastic cellular automaton of this sort? What I find most attractive about the analogy is that it pictures historical and social processes as simultaneously characterized by (1) local predictability and regularity, given information about some local domain, and (2) global unpredictability, chaos, and history dependence.[10]

The global unpredictability part is easy to see based on the discussion above. Even if one could know the whole pattern at time t and if one knew the full set of transition rules—by analogy, more than any social scientist could possibly aspire to—it would be impossible to predict whether a war (a red ring) would appear in a particular place at time $t+50$. Will there be a war in Europe in the year 2053? What will the U.S. economic growth rate be next year, or two months from now? The inflation rate, the Dow Jones, the murder rate, the level of "consumer confidence" or societal alienation? Insofar as all such variables are determined by the actions of myriad agents making myriad choices in response to "local" conditions, and whose choices feed back to each other in highly complex ways over time, these variables are determined by cellular-automatonlike processes. So if the analogy holds, we can forget the goal of making highly accurate point predictions about such things.

According to the analogy, however, this unpredictability and hopelessness of point predictions follow from local-level rules and mechanisms that may be highly regular and predictable. Take, for example, a human life.[11] Locally, my life is highly predictable and so are those around me. I can predict with a high degree of confidence who will come to work at the department tomorrow, for instance. Every day, I and everyone else make thousands of extremely accurate predictions about other peoples' choices and behavior. When we drive, we regularly stake our lives on the accuracy of these predictions. Beyond our immediate environs, we can often make quite sharp predictions about international political events. In late August and early September 1994, many were able to predict with a high degree of confidence that the Clinton administration would not completely "back down" in its confrontation with the Haitian military leadership—Clinton could not have done so after having created all the audience costs he would have suffered for not following through in some way on his massive display of force.

At a longer range, however, lives are highly unpredictable and subject to enormous variation due to very small and often random factors. Children, whose production involves a natural lottery, are a prime example. So are the often accidental circumstances that tip a person towards one career or another. I can predict a few things about what my life will look like in thirty years (if I last that long), but many others are globally unpredictable in the sense discussed above. At a "local" level one can give coherent and convincing causal explanations for life patterns and choices, but for a whole life from start to finish the best we can do is often narrative ("this happened, then this happened, . . .").

The analogy suggests a program for social scientists: Discover and explain the mechanisms, or local transition rules, that make things somewhat predictable at "local" levels.[12] For more global or "macro" levels, the analogy would suggest that all we can do is describe or narrate how various essen-

ization for all we can possibly know about what happens in an individual's head"; or (3) the stochastic element reflects the social scientist observer's lack of information about facts or causal mechanisms at some more micro level.

[10] I stress that I do not think the comparison of historical processes to cellular automata is anything more than an analogy. I can think of virtually no real social process that could be naturally or constructively modeled in this way, with the possible exception of housing and segregation patterns. And I certainly do not mean to say that all of history could be constructively modeled by an automaton. This is just an analogy. Further, it is meant as a loose analogy, in that I do not want to commit to saying that a "cell" is analogous to a state, a group, a person, a neuron, or a particular location in space. The level of application of the analogy is left open.

[11] Reisch (1991) uses the same example, although he draws out quite different implications.

[12] I view game-theoretic models as tools for discovering, exploring, and clarifying arguments about local (versus global) mechanisms of this sort. The idea of social science as seeking to discover and understand local mechanisms and how they work has been consistently developed by Elster (1989; 1993).

tially accidental conjunctions of mechanisms selected one historical trajectory from many other possible ones, which may not even be imaginable in any useful detail.[13] Thus there may be grounds for both the social scientist's view that one can discover meaningful causal patterns and relations in the social world and for a view typically held by historians, that the empirical evidence of history reveals tremendous contingency and essentially unique sequences of events.[14]

Before developing this analogy in regard to counterfactuals and causal arguments about social events, I wish to note two other general aspects that I find appealing and that speak to some issues that often arise in discussing counterfactuals.

First, while the evolution of the patterns on the screen might be highly sensitive to some small events—for example, whether a particular cell happens to flip yellow or green in a given period—this would not necessarily be true of all small events. It could be that in a given period, some cells are positioned in the whole pattern in such a way that their color is highly influential for future states, while others are positioned so that it does not matter at all whether yellow or green occurs. In a deterministic automaton, for instance, two distinct patterns can have the same successor pattern, so that either one yields the same future sequence.

My intuition is that this is also true of historical processes. Whether Khrushchev wore a blue shirt rather than a white shirt on October 25, 1962, probably made no difference whatsoever to the resolution of the Cuban missile crisis. But I can imagine that whether the day was sunny or cloudy, or incidental things that relatives or other politicians said to him, might conceivably have mattered.[15] Assassinations seem to provide the best class of examples here. For these one can often construct plausible arguments about the world-shaping importance of very "small" events, such as whether Lee Harvey Oswald had his morning coffee or not. The analogy suggests that one might simultaneously agree with the historian who asked rhetorically "Can one seriously believe that if my dog whose name is 'Trailer' had been called 'Tiger' everything else would have been affected?" and with the view that some "small" events have enormous consequences (Hook 1943, 121). Everything need not be connected to everything else, especially in the short term. What some Malaysian farmers were doing in 1960 need have had no

[13] Almond and Genco (1977) criticize "behavioralist" political scientists for treating politics as analogous to clocks rather than clouds; they say the latter is more often a better analogy. In Almond and Genco's language, the point of this paragraph can be expressed as follows: Many little clocks can produce big clouds. Discover and study the clocks, describe the clouds and how they were produced from clocks.

[14] For a fairly typical recent statement of this view see Gaddis (1992).

[15] On the influence of the weather: Saunders (1993) shows that New York Stock Exchange prices are systematically lower on cloudy days in Manhattan.

effect on the occurrence or resolution of the Cuban missile crisis, even if in 300 years the world historical consequences of some very small choices by these farmers will be enormous (for example, by influencing who their myriad descendants are and what they will do).

Second, the automaton analogy admits the possibility of statistical regularities, and thus statistical predictability, at the global level. It could be, for instance, that on average ten red rings appear every hundred periods and that this is quite regular. Thus it is not the case that all sorts of events in a cellular automaton need be locally predictable but globally unpredictable—some phenomena may be globally predictable in the statistical sense but locally unpredictable for an observer lacking factual information or the relevant transition rules. Analogously, the suicide rate for a country may be quite predictable year to year even though it would be impossible for an observer with no information about individuals to predict precisely which ones would commit suicide. Nonetheless, few would call a statistical regularity at a global or macro level an "explanation" by itself. Rather, we usually want some reference to transition rules or mechanisms that give rise to the regularity (for example, Catholics have lower suicide rates than Protestants because they are bound more tightly within an integrated religious community) (Durkheim 1951).[16]

Applying the Analogy

In this section I use the analogy of historical processes to cellular automata to consider the two questions identified in the introduction. First, how can we know with any confidence what would have happened if some factor had not been present? And second, are some factors "illegitimate" as counterfactual antecedents, and under what circumstances?

We can distinguish two types of explanatory tasks. First, a researcher may be interested in discovering or testing a factor that is proposed as a cause of some class of events, such as wars, revolutions, national income across countries, protectionist trade policies, and so on. Second, a researcher might be interested in discovering or testing a factor that is proposed as a cause of a single, particular event. By analogy, these tasks correspond to asking

[16] Peter Woodruff (personal communication) has pointed out another sense in which a deterministic automaton may be considered globally predictable but locally unpredictable: In the deterministic case, one can predict the full (global) pattern in period $t+1$ given knowledge of the global pattern in period t, but one cannot predict the $t+1$ pattern for any ("local") contiguous subset of cells without knowledge of the neighboring cells' colors in the period t. I mean "locally predictable" in the sense that to predict the color of a cell in period $t+1$ one needs to know only the colors of the eight immediate neighbors, while to predict the same cell's color in $t+i$, for example, one needs to know the colors of $(2i + 1)^2 - 1$ neighboring cells in period t (if this does not already encompass the whole automaton).

about either the causes of the appearance of red rings in general or about the appearance of a particular red ring.

There are essentially two approaches one can take. First, one might try to evaluate whether the proposed causal factor is regularly associated with the event to be explained across cases. That is, one tries to learn whether the event always or usually occurs when the causal factor is present, but rarely or never occurs when the factor is not present. This is the essence of Mill's method of difference, the approach at the base of all statistical methods for causal inference. Second, one might try counterfactual argument. Here one begins by taking a single, particular event, and then asks whether the event would not have occurred if the proposed causal factor had been absent but all else had been the same.[17]

It is worth noting that either approach can be used for either explanatory task. One might try to explain a particular event by looking for causal factors that were present in this particular case and which are known (or found) to be causal factors by an examination of frequencies of association across cases. Or one might try to show that, for example, imbalances of power cause war in general, across cases, by arguing counterfactually case by case for a sample of wars. It should also be stressed, however, that even when the counterfactual approach is employed in the hope of testing or making a general argument about cause and effect, it does so by focusing on single, particular cases. In the counterfactual approach, evidence always comes from consideration of particular cases, rather than from blunt regularities of association across multiple cases (Fearon 1991, 175–76).

Using the analogy to cellular automata, exactly how does the counterfactual approach work? Suppose we are trying to explain the appearance of a particular red ring pattern in a particular period t. The counterfactual approach proceeds by arguing that the colors of some set of cells w, x, y, . . . in previous periods "caused" the red ring, on the grounds that if these cells had assumed different colors (perhaps specified), the red ring would not have appeared. This claim is then evaluated or rendered plausible by referring to our knowledge of, or beliefs about, the relevant transition rules (the "local" mechanisms), which may take the form of theories or statistical generalizations from observation of past occurrences of red rings. If we know or have some confidence in the relevant transition rules, we may then be able to deduce that given different "initial" colors for cells w, x, y, . . . , the red ring would not have appeared. The rules plus initial conditions imply that something else would have appeared, or that the status quo would have continued.

So, as many have argued, in the counterfactual approach the test of an empirical claim (A caused B) makes crucial and deductive use of some set of theories or statistical generalizations about how the world works. Our grounds for considering these empirically valid must derive either from deductive counterfactual analysis of other particular cases or from the inductive method of comparing across sets of cases. Presumably, at some point the theories or generalizations must be based on the method of induction. For example, if we observed the operation of a cellular automaton and initially had no idea what any of the transition rules were, we could not begin to figure them out by using the counterfactual approach. Instead, we would have to draw inferences by looking for regularities across cases.

How Can We Know What Would Have Happened If . . .

So it appears that the only possible answer to the question "How can we know with any confidence what would have happened if the proposed causal factor had been absent?" is: By deduction using transition rules (local mechanisms or theories sometimes understood only as statistical generalizations) in which we already have some confidence.[18]

If the analogy of historical processes to cellular automata is reasonable, however, then this answer implies a bleak assessment of the prospects for knowing "what would have happened" in many counterfactual analyses. It is in the nature of automatonlike processes that one may not be able to predict what will happen several steps ahead without explicitly "running" the system. Attempting to proceed by deduction from rules plus initial conditions will take one only a very short distance forward before complexity overwhelms the effort. Moreover, this can be true *no matter how good our knowledge of the transition rules or theories used to draw out the consequences of a particular counterfactual antecedent*. Indeed, even in the impossible case of perfect knowledge of all "social science transition rules," we would not be able to say what would have happened without access to a

[17] The underlying logic is the same in either case (Fearon 1991). The question is whether comparison cases are found in actual or possible worlds.

Political scientists who rely on case studies frequently take Mill's method of agreement as another valid approach, and use it to justify research designs where all cases had the same outcome on the dependent variable. As Mill (1900, 286ff) noted, however, the so-called method of agreement *does not work* when there is more than one cause of the phenomenon in question. Because there is no a priori way of knowing how many causes a phenomenon has, and because in political science problems we invariably think there are multiple causes, the method of agreement becomes in effect a rhetorical device that rationalizes bad research designs. Consider the following example. Suppose we want to learn the causes of fatal car accidents, and we observe two cases in which drivers went off the road and into a tree late at night. In one of the two cases the driver was drunk. Using the method of agreement, we conclude that alcohol cannot be a cause of fatal car accidents. This is obviously nonsense.

[18] The Stalnaker-Lewis "possible worlds" approach may represent another feasible answer, or may simply restate this argument in different language; I am not qualified to judge. See Stalnaker (1968) and Lewis (1973, 57, 65–72).

model that was just as complex as the social world itself. In other words, we would need a map just as large as the terrain it described.

From this perspective, the problem of determining "what would have happened if . . ." is fundamentally similar to the problem of forecasting specific political and economic events. If we had the ability to forecast with a high degree of confidence whether and where, for example, the United States will be at war in two years' time, or what GNP growth will be, then we would also have the ability to delineate counterfactual scenarios on such matters with a high degree of confidence. And wherever we are unable to forecast political and social events with confidence, there we will also be unable to develop counterfactuals with much plausibility. The ability to forecast—that is, to make point predictions—entails an ability to use causal theories plus knowledge of initial conditions to spell out a narrative, a chain or succession of events. Often this is precisely what is needed to render a counterfactual claim plausible. But if the automaton analogy holds, then the only successful point predictions social scientists will be making will be quite local or short-range. One might successfully use theories about decision making and strategy to predict the Gulf War in September 1990, but no theories could reliably predict in 1989 the war that occurred in 1991 (let alone predict it in 1945). If this line of argument is correct, then detailed counterfactual scenarios will have a chance at being rendered plausible only if the proposed causes are temporally and, in some sense, spatially quite close to the consequents.[19]

If we think specifically about the analogy of historical processes to *stochastic* cellular automata, a new issue arises that yields new problems. In the paradigm of the perfect thought experiment, we try to imagine the world without the proposed causal factor *A* but with all else the same as in the world that actually occurred.[20] But if we think that historical processes have random components, then the meaning of "all else the same" is not clear. Do we try to sketch the counterfactual scenario so as to be as close as possible to what actually did happen, or do we presume that what did happen was just one of many possible paths, and not necessarily the most likely one? For instance, consider the claim that "if I had not gotten stuck in a

traffic jam, I would have arrived on time." Should we evaluate this claim by using information about how other drivers actually did behave, for example, on the way to the freeway where the jam was, or should we imagine all counterfactual worlds where they might have acted differently due to random variation? I do not see any obviously correct answer here.[21]

As a more applied example, consider the counterfactual claim that if Khrushchev had "stood firm" in the Cuban missile crisis for two more days, Kennedy would have offered a more attractive compromise that Khrushchev would have accepted; thus the risk of nuclear war was really not very great in the crisis. One might support this claim by citing evidence that Kennedy was indeed secretly planning to make a more generous and public compromise offer, and evidence that he was not at all sanguine about the prospects for a disarming attack on Cuba (Blight and Welch 1989, 83–84, 173–74). But in offering this scenario we assume that (1) in the two-day interval that did not actually occur, other events would not have happened that might have influenced Kennedy's or Khrushchev's decision calculus; and (2) events that did occur in the actual world would also have occurred in the counterfactual world. We implicitly extend the pattern of the previous actual days forward into the counterfactual two days, and we also presume that events that occurred in the actual two days would have occurred in the counterfactual two days. How warranted is this? Perhaps if Khrushchev had not offered to settle when he did, there might have been a dispute between Cuban and Russian officers in Cuba that U.S. leaders would have interpreted as seriously threatening and calling for a preemptive strike. And suppose that in the actual world a Cuban and Russian officer happened to have a sharp personal conflict the day after Khrushchev signaled his willingness to withdraw the missiles; does this conflict also occur in the counterfactual world where Khrushchev "stands firm"?

The point is that if we think historical processes evolve stochastically—or stochastically for all we can possibly know—then counterfactual antecedents do not imply determinate paths for counterfactual scenarios, as we often assume in sketching them. Rather, changing a factor counterfactually implies a probability distribution over many counterfactual paths, in which the evidence provided by the actual world that did occur may not be relevant for saying what would have occurred in unprecedented counterfactual situations.

[19] For arguments on the difficulty of forecasting in international politics see Jervis (1991). I should note that making point predictions is not the be-all and end-all of social science. Social scientists seem to do better at making comparative statics analyses of ongoing regularities of behavior or institutions—for example, at explaining why intra-alliance politics takes one form under bipolarity and another under multipolarity.

[20] Mill's (1900, 256) definition of the method of difference suggests just such a perfect experiment: "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."

[21] Using the analogy, the problem may be expressed as follows. We ask what would have happened in period t if cell x had been a different color in time t-3 (for instance). Suppose we know that in the actual unfolding that occurred, in time t-2 a nearby cell z happened to flip green instead of yellow, and that this bore on the outcome in question. Does "all else the same" mean that our counterfactual scenario should leave z green (assuming that the counterfactual change of cell x does not affect z one way or the other), or should we allow for the possibility that yellow might have occurred?

This raises a difficult question about the meaning of "all else equal" in a counterfactual scenario, and adds to the difficulties involved in knowing how plausible our counterfactual assertions are.[22]

To summarize, we will be able to judge the plausibility of and have confidence in counterfactual arguments in precisely those domains where we are confident about making forecasts and predictions. Thus, I may have considerable confidence in the claim that "if I had not gotten stuck in the traffic jam, I would very probably have arrived at class on time." This is plausible because it is asserted in a highly predictable, "local" domain where the causal mechanisms are well-understood and relevant intervening factors are few. Accidents may happen, but I know them to be rare from past experience.[23] By contrast, the claim that "if Gorbachev had lost the succession struggle in 1985 the Soviet Union would not have disintegrated before 1992" is asserted in a domain where causal mechanisms are less well understood and, moreover, *even if they were well-understood* a forecast such as this one might be impossible because it is too "long range."[24] The greater the number of steps and the more time passes, the more automatonlike chaos will intervene.

So what do we do with counterfactuals like this one about Gorbachev? These are often precisely the counterfactuals in which we have the most interest. If there is no way even in principle that one could say precisely what would have happened in 1992 if Gorbachev had lost the succession struggle in 1985, then what is the value of making such claims? Or consider the claim that if nuclear weapons had not been invented, the post-War world would still have seen major-power peace. There is absolutely no way of anticipating a forty-five-year path of counterfactual events with any confidence, so what empirical evidence relevant to assessing whether nuclear weapons caused the "long peace" can be gathered by exploring such a scenario?

[22] This is exactly the problem posed by Robyn Dawes in his "betting on snake eyes" example (Commentary 3), where he insists that "all else equal" should allow for stochastic variation. Likewise, King, Keohane, and Verba (1994, 76–85) give a definition of causality that comes down on the side of always allowing for counterfactual stochastic variation. It seems to me that this may make sense if, in thinking counterfactually about one particular case, one is trying to extract a generalizable "causal effect." But if one is asking "what would have happened *in this particular case* if such-and-such had been different?" then our intuition demands that we try to hold all other things equal as we knew they actually happened.

[23] Alternatively, one might say that we can assess counterfactual claims in precisely those areas where we can also use the method of difference on multiple actual cases. "I would have arrived on time" may be plausible in light of the comparison to many other typical driving days I have experienced.

[24] Arguably the relevant "local-level" mechanisms *are* well understood—we rarely have much difficulty explaining why any particular politician made a particular choice at a particular time, given enough information. It is in aggregating all the various decisions and situations that complexity overwhelms.

My sense is that counterfactual claims such as these two rarely or never act as independent empirical tests or sources of empirical evidence. Rather, they are typically rhetorical devices—"spotlights" in Mark Turner's (Commentary 1) terms—that at best point the reader towards bits of evidence relevant to assessing causal claims that may be somewhat different from the one suggested by the counterfactual. We cannot say with any confidence what would have happened in 1992 if Gorbachev had lost out in 1985, but the counterfactual claim may serve to direct our attention to the political programs being offered by Gorbachev's rivals, to the sources of their support, and to a comparison of their strength to the strength of Gorbachev and his camp. The counterfactual claim is thus a rhetorical flourish directing attention to another, more "local" and empirically assessable claim: that the political program of Gorbachev's closest rival was quite different, and he would have had a good chance of implementing it, at least in the short run. Thus, had Gorbachev lost the succession struggle, the Soviet Union would not have started down the specific path that we saw lead to disintegration. Where it would have led by 1992 will always be difficult or impossible to say.

Similarly, while we cannot hope to describe a post-War world without nuclear weapons in any convincing detail, making the rhetorical claim directs our attention to the somewhat more assessable question of what factors prevented war between the United States and the Soviet Union in the actual world that did occur. One can ask, for example, how large a role nuclear weapons played in actual superpower crises.[25] If one answers "very little"—most or all crises would have worked out the same way even without nuclear weapons—then this would count somewhat against the view that nuclear weapons caused the long peace. It could hardly be decisive, of course, because nuclear weapons might still have been responsible for moderating superpower behavior so that the crises that did occur were fewer and less serious than would otherwise have been the case. On the other hand, if one finds evidence that concerns about nuclear war were a significant factor moderating U.S. and/or Soviet behavior in the actual crises, then this would count in favor of the claim. Again, the evidence would not be decisive, because it could be that nuclear weapons were themselves a principal source of war-threatening friction, and so gave rise to dangerous crises that would not have otherwise occurred. To some extent this possibility might be assessed by asking about the principle interests in dispute in the several major crises that occurred, which would lead in turn to counterfactual speculation at a more "local" level. For example, would a crisis over Cuba have occurred at all but for nuclear weapons? The immediate cause of the missile crisis was, of course, the introduction of nuclear missiles, but absent these a

[25] Some relevant evidence is presented in Betts (1987).

war-threatening crisis might have arisen if the United States in a non-nuclear world had tried harder to overthrow Castro. Perhaps in this world the Soviets would have viewed the option to take Berlin more favorably.

My general point is that the value of counterfactual claims often does not lie in the possibility that empirical evidence can be adduced by explicitly exploring the counterfactual scenario. Complexity means that for all but the most local-level claims, we simply cannot say with any confidence what would have happened. Instead, unassessable counterfactuals typically act as rhetorical devices or "spotlights" that direct us to look at other, more local sorts of evidence relevant to assessing related causal claims.

Are Some Counterfactual Antecedents More Legitimate Than Others?

Although it may be difficult or impossible to describe in any detail the world that would have followed if causal factor *A* had not occurred, it is frequently easy to support a weaker claim that whatever would have happened, it would not have been effect *B*. For example, in an article attempting to explain the failure of the August 1991 coup against Gorbachev, Stephen Meyer (1991, 5) writes that "there can be little doubt that had the military establishment actively supported the putsch the outcome would have been far different." He makes no effort to say exactly what would have happened—presumably it would be difficult to predict much beyond the immediate success of the plotters and the crushing of demonstrations. But he asserts with justifiable confidence that the coup would not have collapsed as it did.[26]

The automaton analogy suggests why this is the case. Consider the problem of explaining the appearance of a red ring in period *t*. The number of cells in prior periods that, had they been different, would have caused *B* not to occur may be tremendously large, and we may be able to assert this with confidence even if we cannot say exactly what would have occurred. Indeed, the farther back we go the more cells will matter "causally" in this sense. Pascal's example concerning Cleopatra is of this type: We cannot know exactly what would have followed if Anthony had not fought a war on her behalf, but we can be reasonably confident that history would have taken a very different course. Thus we face the prospect of finding the claim "if Cleopatra's nose had been a different length, World War I would not have occurred" more plausible and defensible than "if Lord Grey had sent a stronger signal of Britain's willingness to fight, World War I would not have

[26] Max Weber's (1949, 164–88) example of an important counterfactual argument in history also takes this form: Eduard Meyer claimed that if the Athenians had lost at Marathon to Persia, the West would not have developed with many of the features that seem to distinguish it from the East. Meyer does not attempt to fantasize about the specific course that would have been followed if the Persians had won, only to show that evidence suggests that from this starting point the trajectory would have been very different.

occurred."[27] Ironically, "butterfly effect" counterfactuals of the Cleopatra's-nose type may be among the most defensible and plausible even as they seem intuitively wrong, or not what we are after in seeking causes.

There are other cases of counterfactual claims that seem highly plausible but whose antecedents strike us as wrong or "illegitimate." The claim that "if Napoleon had had Stealth bombers at Waterloo, he would not have been defeated" may be plausible in that if we grant the antecedent, the consequent might very well follow (that is, we could make a strong argument for it). But we are very reluctant to grant the antecedent, which seems illegitimate because we think that Napoleon "could not possibly have had" Stealth bombers.

Ruling out antecedents of this sort puts us on a slippery slope, however. To do so consistently we need criteria for distinguishing between legitimate and illegitimate antecedents. In effect, we need criteria for deciding what factors or events should be considered as possible causes of a phenomenon, and what factors should not be considered as causes. Such criteria are not easy to produce.

The standard suggestion is that the counterfactual antecedent must have been "actually" or "objectively" possible.[28] Even assuming that we can say what it means for something to be "actually possible" at a given time, this criterion leads to substantial difficulties. Most importantly, in both ordinary language and in social science research we often call factors "causes" that do not meet this criterion. We may say that a person's death was caused by old age even if it was not actually possible that the person had been younger. Or consider a study of voting behavior that uses regression analysis to assess the causal impact of factors such as race, religion, party identification, and income on vote choice. The study may conclude from the data that race has a significant causal impact on individual voting behavior, even though it is not actually possible that any individual survey respondent could have been a different race. The same applies for large-*N* studies of deterrence, which may find that a balance of forces causes deterrence even if it was not "objectively possible" that the balance could have been much different than it was for particular cases in the sample.[29]

[27] If the reader has doubts about Pascal's specific example, then substitute any event in ancient times that can be plausibly argued to have had a major impact on the course of all subsequent history (e.g., the Battle of Marathon).

[28] The argument is most powerfully and emphatically expressed by Elster (1978, 185). Barry (1980) argued against this view in his review of Elster's book. For other authors taking the "counterfactuals must have been actually possible" position see Hawthorn (1991, 158–59) and Tetlock and Belkin on "minimal-rewrite rules" (Chapter 1). It is worth noting that this criterion does not help for examples of the Cleopatra's nose sort—butterfly-effect causes may be "actually possible" and so legitimate in this sense.

[29] In their discussion of causality, King, Keohane, and Verba (1994, 78) endorse the view that the counterfactuals employed in either large- or small-*N* research should be "reasonable and it should be possible for the counterfactual event to have occurred under precisely stated cir-

There are two problems of "legitimacy" here. First, are butterfly-effect causes—which include more than fanciful examples like that given by Pascal—legitimate as counterfactual antecedents? Second, in using counterfactuals to assess proposed causes, can we vary factors that either "had to be as they were" or "could not possibly have been"? I will suggest that there are no hard and fast, "scientific" answers to these questions because they are really questions about what our intuition will accept as a cause, and to answer them we have to ask about what we mean by the word in different contexts. I will consider each problem in turn.

BUTTERFLY-EFFECT CAUSES

When we explain why an event *B* occurred, we explain why *B* occurred *rather than* some other alternative or set of alternatives. As Alan Garfinkel (1981, chapter 1) has argued, explanation always takes place relative to a "contrast space" of alternatives, and how this space is implicitly imagined strongly influences what a satisfactory explanation will be. For example, consider the question "Why did the Soviet Union disintegrate?" A political scientist might reject the explanation that "a modern economy simply cannot work with central planning rather than markets" on the grounds that this does not explain why the Soviet Union fell apart when it did. But the central planning explanation might be perfectly acceptable if the question is asking why the Soviet Union disintegrated at all rather than surviving indefinitely. We imagine a different contrast space when we ask "Why did the Soviet Union disintegrate in 1991 rather than at some other time?" and this affects what is acceptable as an answer.[30] Similarly, the claim that "ancient hatreds are a cause of the war in Bosnia" may be a respectable part of an answer to the question "Why is there an ethnic war in Bosnia *rather than* in France, Britain, the United States, etc.?", but it would not answer the question "Why is there a war in Bosnia now rather than at other times?" Particularly when asking about the causes of a specific event, social scientists frequently fail to specify the contrast space, and this often leaves crucial ambiguity about what question is being asked.[31]

Invariably, however, when we try to explain why some event *B* occurred, we implicitly imagine a contrast space in which *B* is absent *and the rest of the world is similar to the world in which B* is present. To show that *A* caused *B*, the relevant counterfactual to make plausible is not "if *A* had not occurred, *B* would not have occurred," because this admits a contrast space that includes worlds where *B* did not occur and the rest of the world was entirely dissimilar to the world that did occur. Instead, the relevant counterfactual should be, "if *A* had not occurred, *B* would not have occurred and the world would be otherwise similar." The goal is to explain the presence or absence of *B* against a fixed, actual "background," rather than *B* in the context of all conceivable backgrounds.

cumstances." But this would rule out much of what the authors take to be paradigmatically good social science! Authors estimating causal effects using regression analysis on nonexperimental data *never* ask whether it would have been actually possible for each case in the sample to have assumed different values on the independent variables.

[30] In the first case, the contrast space is {Soviet Union exists indefinitely, Soviet Union collapses at some time}. In the second, the space is {Soviet Union collapses in 1991, Soviet Union collapses at some other time}.

[31] Indeed, the principle benefit of explicitly specifying the counterfactual implicit in a causal claim may be that doing so forces the researcher to articulate the contrast space and thus exactly what he or she is trying to explain (cf. Fearon 1991, 194).

Butterfly-effect "causes" seem intuitively peculiar for precisely this reason: They admit a bizarre contrast space, one different from what is implicitly assumed in our idea of explanation. World War I may not have occurred if Cleopatra's nose had been different, but the butterfly-logic behind this claim also implies that all aspects of the world in 1914 would have looked tremendously different. Because the implicit assumption was that we were looking for causes of the occurrence of World War I in a world that otherwise looked as it did in 1914, Cleopatra's nose does not work as a "cause" in our standard (intuitive) sense.[32]

Thinking in terms of the automaton analogy, we can acknowledge that an enormous number of prior events (cell colors in previous periods) may have been necessary to determine the appearance of a particular red ring at a particular time, while at the same time denying that the vast majority of these were "causes" of the red ring in the normal sense of the word. For thousands and thousands of prior events it may be true that if the color of the cell had been different, the red ring would not have appeared when and

[32] In my earlier paper, I proposed a somewhat different resolution for the butterfly-effect problem: Cleopatra's nose length was not a "cause" of World War I because the probability of the war occurring conditional on a different nose length was no different from the probability of the war occurring conditional on the nose being as it was (i.e., almost zero in both cases) (Fearon 1991, 191). King, Keohane, and Verba's (1994, 82) definition of causality works the same way: "The causal effect is the difference between the systematic component of observations made when the explanatory variable takes one value and the systematic component of [here, counterfactual] comparable observations made when the explanatory variable takes another [counterfactual] value." Thus, if the independent variable is nose length, there is almost no covariation with a dummy variable for the occurrence of World War I. In the thought experiment of many hypothetical "runs" of the world with different nose lengths, the war virtually never occurs, so its "causal effect" is judged to be almost zero.

Although they work for this example, conditional probability definitions seem to have problems with others. For example, consider the proposition that the presence of oxygen in the atmosphere was a cause of World War I. It might be true that over many hypothetical "runs" beginning in July 1914 the war never occurs when there is no oxygen and that war almost always occurs when there is. So by a conditional probability definition, oxygen counts as having a large causal effect on the war. More in accord with intuition, it does not in the definition suggested above, because the world is not "otherwise similar" if oxygen is subtracted.

where it did. But for a much smaller number would this proposition be true and it would also be true that the rest of the pattern on the screen would be similar to what did occur with the red ring. I would argue that only events of the latter type meet our intuitive notion of cause.[33]

This suggestion is problematic in that it relies on an unelaborated metric for judging similarity across worlds, and also assumes that we can assess moderately well whether the rest of the world would be sufficiently "similar" following a change of the proposed causal factor. At a certain temporal or spatial range, these distinctions and forecasts are hard to make. Consider the claim that social Darwinism was a cause of World War I. Suppose Darwin had died young, and that a theory like Darwin's did not develop until sometime in the mid-twentieth century. Without Darwin and Darwinism in the 1860s, it is entirely plausible that what we call World War I would not have occurred, although perhaps some other big war might have. The "rest of the world" would look in many respects like the Europe that did exist in 1914, but it might be very different in other respects (for example, more uniformly liberal or socialist, and less imperialist). By my criteria, it is hard to say whether social Darwinism ought to count as even a possible cause of the war.

How one judges whether the world would have been "otherwise similar" depends in large part on how narrowly one defines the event being explained. Using the example above, does "World War I" mean (1) a war between the five European great powers that begins in 1914; (2) a war between at least four great powers beginning in the period 1890–1920; or something in between? If Darwin had not occurred, it seems entirely likely that (1) would not have occurred, for butterfly-effect reasons. But (2) might or might not have occurred. So if "otherwise similar" means a world as narrowly defined as that described in (1), then the criterion given above rules out Darwin as a cause. By contrast, if we take a broader class of "World War I's" and late-nineteenth century Europes, then Darwinism might or might not be judged a cause, depending on how much one thinks social Darwinist thought contributed to the practical views of European leaders on war.

This discussion raises a more general and very important point: What we will accept as possible causes of an event (or class of events) depends crucially on the level of detail with which we specify the event. As Alan Garfinkel (1981, 28–32) has argued, when we try to explain some particular occurrence we always implicitly imagine a class of events that would qualify. For example, it would not disqualify an explanation of the occurrence of the French Revolution if the explanation did not account for Robespierre's choice of clothes on a particular day, even if this does make up part of the specific occurrence that was the French Revolution. Rather, by "French Revolution" we implicitly have in mind a large class of occurrences, all of which we would accept as "essentially equivalent" to the revolution that did occur, for the purposes of the explanation. Depending on the explanatory focus, an explanation of World War I might have in mind an equivalence class that included only wars begun in the fall of 1914 over a dispute between Austria and Russia in the Balkans, or it might be broader, accepting any great power war in a twenty-year interval. What one will accept, or even imagine, as causes will of course depend on how broadly the "equivalence class" is defined.[34]

Thus the assassination of Archduke Ferdinand might be reasonably judged a cause of World War I, if the event "World War I" means "a general war among the European great powers beginning in the fall of 1914 over a dispute in the Balkans." But it might not be a cause of a World War I if the event to be explained is less narrowly drawn, meaning, for example, "a general war between the Triple Alliance and Triple Entente beginning sometime between 1910 and 1920." In addition to often failing to specify the relevant counterfactual claims, scholars making causal claims about international politics often fail to specify what would qualify as the event they are trying to explain. Because this obviously affects the truth or falsity of counterfactual arguments used to support or disconfirm possible causes, it is critically important to be careful about it.[35]

CAUSES THAT "HAD TO BE AS THEY WERE" OR "COULD NOT POSSIBLY HAVE BEEN"

As noted above, many writers on the use of counterfactuals in social science have argued that some counterfactual antecedents are more "legitimate" than others, and thus that some counterfactual propositions should not be entertained in seeking to learn or assess the causes of an event. Examples range from the relatively fanciful (but still problematic) Stealth bomber sort to

[33] When historians and social scientists argue that each event is historically unique and produced by an infinite stream of particular prior "causes" as far back as one wishes to go, they are employing a "butterfly-effect" notion of cause rather than our ordinary language one. See, for an example, Weber's (1949) discussion, which makes several references to particular events being caused by an "infinite succession" of prior events and circumstances.

[34] See A. Garfinkel (1981, chapter 1) for the term "equivalence class" and a discussion. Another implication of Garfinkel's insight is that historians and social scientists who claim to be explaining "particular" or "unique" events are never really doing this, in a literal sense. Instead, they must have in mind a class of hypothetical events that are all essentially equivalent as far as their explanatory purpose is concerned.

[35] Thus, a principal reason for the methodological rule that the antecedent and consequent in a counterfactual need to be clearly specified is that the precise definition of the consequent crucially determines what may have caused it (Chapter 1, Tetlock and Belkin; Fearon 1991).

more difficult questions concerning whether certain decisions by state leaders were "actually possible" or not at a given time.[36]

Criteria for deciding the legitimacy of counterfactual antecedents are really criteria for saying what should or should not be considered as a cause of the consequent in question. As I suggested above, the standard suggestion that counterfactual antecedents are "illegitimate" if we have theories and arguments implying that they were objectively impossible cannot make sense of our usage of "cause" either in ordinary language or in typical social science practice. The automaton analogy suggests a distinction that, I believe, is less problematic and generates more insight.

For a stochastic cellular automaton, we explain the appearance of a particular red ring by arguing that if some set of cells *x*, *y*, *z*, . . . in previous periods had taken different colors, the ring would not have appeared and the overall pattern would have been otherwise similar. Beyond this, I put no constraints on how we propose to vary counterfactually the colors of cells *x*, *y*, *z*, and so on. There are two possibilities here. We can either restrict ourselves to color changes that actually could have occurred given the stochastic transition rules, or we can imagine ourselves intervening and changing colors of cells as we please, paying no special attention to the transition rules. In the first approach, we say that if cell *x* had flipped black rather than yellow, which it actually might have done under the rules, the red ring would not have appeared. In the second approach, we say that if cell *x* had flipped blue rather than yellow—which was impossible under the rules—the red ring would not have occurred.

I will call causes in the first approach *conceivable causes*, because they could conceivably have happened according to the "rules of the game" as we understand them. By contrast, I will call causes in the second approach *miracle causes*, because we imagine their counterfactual occurrence as resulting from an intervention from outside the system (the hand of God, as it were).

In ordinary language and in social science practice we rely on both types of cause in different contexts. I doubt that either one can be justified as a uniformly correct notion of cause, considering that explanatory purpose and context seem to determine which one we adopt. Conceivable and miracle causes are supplied in response to different sorts of questions.

More specifically, when we try to give or assess the causes of a particular event, such as Napoleon's defeat at Waterloo, the end of the Cold War, or the collapse of the Soviet Union, we are sometimes asking for conceivable causes and sometimes for miracle causes. In some contexts and for some authors, the question "Why did the Soviet Union collapse?" may be asking "What about the world could actually have been different and led to continuance of the Soviet Union?" Or the intention of the question may be less restrictive. The author may intend to answer "What changes in the world, whether actually possible or not, would have prevented the Soviet Union from disintegration?" Just as there are two different ways to imagine cell colors in a cellular automaton being different, there are two different ways of imagining making counterfactual changes in the thought experiments we use to argue causality.[37]

In what contexts do we expect one type of cause rather than the other? I do not think there are any very sharp rules here, but at least two rough generalizations can be offered. First, when we are treating the specific event to be explained as an instance of a class of events, we are generally quite ready to accept miracle causes. For example, in regression analysis and other statistical means of testing causal hypotheses, one assumes that if any particular case in the sample had taken a different value on one of the independent variables, the dependent variable would have differed by a systematic component that is the same across cases plus a random component. One never even contemplates whether it would have been actually, historically possible for any particular case to have assumed different values on the independent variables. Thus, in research of this sort that seeks causes of recurrent events rather than particular events, there is nothing peculiar about statements such as "if John Smith had been black, he would have been 30 percent more likely to have voted for Clinton than he actually was." By contrast, if we were asking the question "What could conceivably have been different and would have led John Smith to vote for Clinton?" it would seem absurd to use Smith's race as a cause, or to vary Smith's race counterfactually.[38]

Some recent work on deterrence provides a more dramatic example, one fundamentally similar to the supposedly ridiculous Stealth bomber case. In order to assess the causes of successful deterrence in a certain class of international disputes, Paul Huth and Bruce Russett collected data on fifty-eight interstate crises in the period 1885–1983. One of the possible causes they wished to evaluate was nuclear weapons—does the possession of these weapons make it more likely that efforts by a defending state to deter an attack on a smaller protégé by a challenging state will succeed? For each case in the sample, the independent variable was coded "1" if the defender had nuclear weapons, and "0" otherwise. Using a probit model, they then

[36] See, for example, Hawthorn (1991, chapter 3); Breslauer (Chapter 3).

[37] When we use miracle causes in an explanation, we seem implicitly to have in mind the idea of an experiment. In true experiments, the experimenter acts literally as the "hand of God" that intervenes from outside, assigning causes to cases.

[38] In this example, race "explains" a person's vote in very much the sense of Hempel's covering-law model: John Smith's vote is explained by subsuming this case under the "lawlike" principle that whites are more likely than blacks to vote Republican. The covering-law model may be particularly friendly to miracle causes, while narrative or genealogical models of explanation are more friendly to conceivable causes.

estimated an average effect of nuclear weapons on the probability of successful deterrence (along with the effects of other independent variables).[39]

The model thus produces estimates of what would have (probably) happened if, for example, Britain had had nuclear weapons during the July crisis of 1914! But Britain "could not conceivably have had" nuclear weapons in 1914, just as Napoleon could not conceivably have had a Stealth bomber. Does this make inclusion of nuclear weapons in the model "illegitimate"? I think the answer should be no, because what Huth and Russett are doing is only an extreme instance of a sort of inductive reasoning we practice all the time. In this form of empirical assessment, evidence comes from regularities of association across cases, and "cases" are understood not as historical particulars, but rather as ahistorical configurations on independent variables. In the regression analysis, the case of "Britain-Belgium-Germany 1914" is just a list of values of the independent variables being assessed, and is implicitly assumed to be the same (absent the random "other causes") as any case that has these values, regardless of when or where it occurred. It is important to realize that this procedure is not bizarre and unusual—we use it all the time in a less formal way when we give explanations. Particularly when we seek to assess causes of a class of events by looking for regularities across cases, we do not worry about whether in each case it was "actually possible" for the proposed cause to have been present or not. For example, when we say that a particular person's death by lung cancer was caused by smoking, we do not worry about whether the person may have had an "addictive personality" or constitution such that he could not actually have quit (or that he smoked in a social environment that supported smoking and was not aware of the link to lung cancer).

So we may be more likely to expect and use miracle causes when trying to explain classes of events than when explaining singular events. This does not mean, however, that miracle causes are never invoked in efforts to give causes of particular events.[40] In fact, this is a common strategy, especially when a researcher is arguing that the particular event in question was "inevitable" and had such-and-such causes. For example, if I argue that World War I was caused by inevitable shifts in the distribution of military and economic power in Europe in the preceding twenty years, then I am invoking a miracle cause, and nothing seems peculiar about this claim. Similarly, the "ultimate" or "underlying" cause of Soviet collapse is often given as a factor that, it is assumed, could not have been different—the supposedly inherent, inescapable inefficiency of Soviet-style economic planning.[41]

Nonetheless, the more fine-grained a researcher's effort to give causes of a particular event, the more likely that he or she will tend to look for conceivable causes. Typically historians focusing on particular sequences of events and political scientists contemplating counterfactual scenarios are asking about conceivable causes. They want to know what about the world could actually have been different and could have led to a different outcome. Part of what is funny about the Stealth bomber example is that it provides an outlandish miracle cause in a context in which intuition wants a conceivable cause—what we really want to know is whether and what could actually have led to a Napoleonic victory. Stronger evidence for this proposition is that virtually every analyst who has written on counterfactual thought experiments in history or social science has argued or accepted without question that it is "illegitimate" to counterfactually change things that "had to be" as they were—this despite the fact that we do it all the time in framing explanations that we take to be valid.

If this generalization holds, then researchers who seek to evaluate their theories using case studies and counterfactual arguments, or who develop their theories by trying to generalize from particular cases, may tend to be biased towards conceivable causes. Further, because conceivable causes are more likely to be specific to each case, these researchers will be biased against finding causes that generalize across cases, which are by and large what social scientists are most interested in. Historians, who frequently dismiss the whole enterprise of finding causes that generalize across cases, may provide the best example of this bias in action. Historians tend to look for conceivable causes and these rarely generalize much.

From a methodological standpoint, then, it may be valuable to keep the distinction between miracle and conceivable causes in mind when one is trying to use counterfactual argument to assess the causes of some particular event or general phenomenon. Counterfactual analysis may bias one towards conceivable causes by implicitly defining the explanatory problem in a certain way (that is, by making the question "How could things actually have been different?"). But since we are often interested in learning what are the factors that regularly produce some outcome, such as war, democracy, or economic growth, then we probably should not rule out miracle causes by methodological fiat.

Simply saying that in making counterfactual arguments scholars should not restrict themselves to conceivable causes (unless, of course, conceivable causes are what they want to learn) is not enough, however. Because for any particular case there may be a huge number of "miracles" that might have

[39] See Huth and Russett (1984) and Huth (1988). For a reanalysis of the data from a different theoretical perspective see Fearon (1994c).

[40] Or that conceivable causes never appear in studies of recurrent phenomena.

[41] The automaton analogy suggests a plausible interpretation of the idea of historical inevitability, which is sometimes viewed as problematic or incoherent. In a stochastic automaton, almost no event is "inevitable" in the sense that, for butterfly-effect reasons, if something had happened differently many periods earlier, the event would not have occurred. However, it may still make sense to say that if the event occurred in period t, it was "inevitable" (or very probable) in period t-i given the nature of the transition rules and conditions in period t-i.

precluded the consequent while leaving the rest of the world otherwise similar, I may have worsened one of the problems noted at the outset of the paper: When one takes a counterfactual approach to hypothesis testing, far too many valid "causes" may appear. There is also the problem of exactly how one imagines the miracle occurring. Exactly how do we picture the British with nuclear weapons, or the British and French with larger militaries in 1935, or nineteenth-century America without railroads? The specific way we imagine the counterfactual antecedent may strongly condition the conclusions we draw from the counterfactual exercise. At least when we restrict ourselves to conceivable causes there are implicit guidelines about how the counterfactual antecedent is to be introduced—as suggested by Tetlock and Belkin, Hawthorn, and others, conceivable causes should be rendered consistently with historical facts, well-established theories and statistical generalizations, and so on. For miracle causes it is not clear what the guidelines, if any, should be.

I see no easy resolution to these problems, and can offer only some incomplete suggestions. First, because miracle causes seem relatively unproblematic in large-*N* research designs, we might try to follow this example when employing miracle causes in counterfactual arguments about particular cases. In the large-*N*, regularity of association approach, the range of what might be called "permissible miracles" is given by the range of outcomes on the dependent variable for all cases in the sample. Thus, we might make it illegitimate to introduce miracle counterfactual antecedents that have not been realized for any other actual case.

Second, some miracles are easier to contemplate than others due to the fact that there are sufficiently many "like" cases that both we and the decision makers involved would have a sense of the meaning and implications of the counterfactual change. For example, it is less problematic to imagine Britain and France with counterfactually strong militaries in 1938 than to try to imagine Britain with nuclear weapons in 1914, even if we are committed to the view that only by a miracle could Britain and France have been stronger than they were. For the case of nuclear weapons, it is almost impossible to imagine how European leaders in 1914 would think about these devices, or how Continental leaders would react to a British announcement and test, and so on. We have only one instance of the invention of nuclear weapons to go by. By contrast, if Neville Chamberlain awoke one morning in 1938 and was told that, due to an extraordinary failure of military accounting, the reserves and air force were in considerably better shape than had formerly been believed, we can imagine how Chamberlain and others might have responded to this knowledge. It was certainly within their comprehension, considering that people at the time explained events in foreign policy by referring to relative military strengths.

Regarding the second problem of how we should imagine the miracle cause being inserted or subtracted in the counterfactual scenario, I can only suggest a Lewislike closeness criterion: introduce the miracle by making as few changes as one can in the actual world. For instance, suppose one thinks that railroad technology had to have been invented around the time it actually was, so that Fogel's counterfactual exercise in *Railroads and American economic growth* considers what I have called a miracle cause. How best to envision the nineteenth-century United States without railroads? One might imagine counterfactually that for some obscure and purely technical reason, railroads were not feasible; perhaps in a "close" counterfactual world there is no way to lay tracks that can endure more than a few trips by heavy locomotives. Or perhaps for obscure reasons Americans are systematically deluded and believe that canals are much more efficient than railroads, so they never really try railroads. With miracle causes the problem is not to suggest a counterfactual that "actually could have happened" but rather to introduce the counterfactual antecedent so as best to capture the sense and intent of the question "What would have happened if . . . ?"

Conclusion

For diverse reasons, many political scientists who study international relations consider small-*N* sets of case studies the best or only feasible way to test causal hypotheses. Moreover, they frequently choose cases that all have the same outcome on the dependent variable or try to explain the occurrence of more or less "one-time," "unique" events. This approach is maintained despite standard and largely uncontested statistical arguments that say such research designs cannot actually test hypotheses. For example, in the conventional, regularity of association approach, selecting on the dependent variable will produce biased estimates of the impact of the independent variables in question, and considering only one case (i.e., one "data point") makes it impossible to draw causal inferences at all.[42]

As I have argued elsewhere (Fearon 1991), reliance on counterfactuals may be a way that users of case studies seek (mainly unconsciously) to increase their *N*. Counterfactual scenarios may provide the controlled comparisons necessary to support causal inferences when researchers restrict themselves to a small number of actual-world cases. If this is so, and considering that scholars employing case studies for the most part are not very explicit about their use of counterfactuals, we should ask if there are meth-

[42] On the problems created by selecting on the dependent variable see Geddes (1990). Doug Dion (1995) has argued that selecting on the dependent variable is a perfectly appropriate research design if the goal is to test for necessary conditions. See also Collier (1995), who suggests several reasons why research designs that select on the dependent variable may be worthwhile and justifiable.

odological guidelines or even a method of counterfactual argument that would help in this kind of work.

The arguments developed in this chapter lead me to be somewhat pessimistic about the possibility and usefulness of any method of counterfactual argument. The chief problem is that its domain would necessarily be very narrow. As the analogy to the cellular automaton suggests, for typical social science problems we will only be able to judge the plausibility of counterfactual arguments for highly "local" situations. That is, we will be able to assess plausibility only where the counterfactuals invoke causal mechanisms and regularities that are well understood and that are considered at a spatial and temporal range small enough that multiple mechanisms do not interact, yielding chaos. As Sidney Hook (1943, 134) put it years ago, "When we draw the line of possible eventuality too far out of the immediate period, the mind staggers under the cumulative weight of the unforeseen." Better theory may push back the limits of what is "too far" a bit. But if the automaton analogy holds, then even a perfect understanding of the local mechanisms would not allow us reasonably to assess the plausibility of important counterfactuals such as "if Gorbachev had lost the succession struggle, the Soviet Union would not have disintegrated by 1992" or "even without nuclear weapons, the post-War world would have been peaceful."

I would thus suggest adding a *proximity criterion* to the list of guidelines proposed by Tetlock and Belkin for assessing the plausibility of counterfactual thought experiments. The criterion may be stated as follows: Consider only thought experiments in which the hypothetical antecedent and consequent are close together in time and are separated by a small number of causal steps. A great many counterfactuals that are, unfortunately, of great interest will be deemed unassessable by this criterion. This does not mean that they should never be posed—such counterfactuals may serve as valuable "spotlights" directing our attention to more local and assessable counterfactuals. Rather, exploring counterfactual claims that fail the proximity criterion is unlikely to yield any very defensible judgment on the causes of the event in question.

The real payoff for carefully specifying social science counterfactuals will probably not be found in some generalizable method of empirical evaluation. There may be occasions when exploring a counterfactual scenario will allow a highly plausible test of a hypothesis, but I doubt the circumstances are very general or that there exists a method of general application.

Rather, the main benefit of being careful about counterfactuals is that doing so forces one to be clearer about, or to "unpack," the nature of the explanatory exercise one is engaged in. Specifying and exploring the counterfactuals implied by a causal claim forces one to be clear about (1) the precise delimitation of the event being explained; (2) the "contrast space" or set of alternative outcomes from which the event that occurred is explained; and (3) the type of causes one is looking for. Regarding (3), I have proposed distinguishing between conceivable and miracle causes, and have suggested that we should only consider counterfactual antecedents that might affect the consequent while leaving the rest of the world otherwise similar.

Exploring counterfactuals opens up a range of difficult and often philosophical questions concerning what we are doing when we try to explain particular or recurrent international political outcomes. A final benefit of thinking about counterfactuals is that doing so brings some of these foundational issues out into the open. Failing to carefully specify the requisite counterfactuals is a way of sweeping such questions and problems under the rug. The more we keep these problems hidden from view, the more our "explanations" will have the character of persuasive rhetoric rather than empirical discovery.