Published by Princeton University Press, 41 William Street,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press, Chichester,
West Sussex

Library of Congress Cataloging-in-Publication Data

Counterfactual thought experiments in world politics : logical,
methodological, and psychological perspectives / edited by Philip E.
Tetlock and Aaron Belkin.
p. cm.
Includes bibliographical references and index.
ISBN 0-691-02792-7 (cl : alk. paper). — ISBN 0-691-02791-9 (pb :
alk. paper)
1. World politics. 2. History—Philosophy. 3. Counterfactuals
(Logic). I. Tetlock, Philip. II. Belkin, Aaron, 1966–
D16.9.C645 1996
901—dc20
96-6396
CIP

This book has been composed in Times Roman

Princeton University Press books are printed on acid-free paper and
meet the guidelines for permanence and durability of the Committee
on Production Guidelines for Book Longevity of the Council
on Library Resources

Printed in the United States of America by Princeton Academic Press

10 9 8 7 6 5 4 3 2 1

10 9 8 7 6 5 4 3 2 1
(Pbk.)

Contents

PART ONE: Counterfactual Inference: Form and Function

PART TWO: Counterfactual Analysis of Particular Events

PART THREE: Counterfactual Analysis of Classes of Events

PART FOUR: Counterfactuals and Game Theory

PART FIVE: Computer and Mental Simulations of Possible Worlds

PART SIX: Commentaries

Contributors

AARON BELKIN
Doctoral candidate, Department of Political Science
University of California, Berkeley

GEORGE W. BRESLAUER
Professor, Department of Political Science
University of California, Berkeley

BRUCE BUENO DE MESQUITA
Senior Fellow, Hoover Institution
Stanford University

LARS-ERIK CEDERMAN
University Lecturer in International Relations, Somerville College
Oxford University

ROBYN M. DAWES
University Professor
Carnegie Mellon University

RONALD J. DEIBERT
Assistant Professor, Department of Political Science
University of Toronto

JAMES D. FEARON
Assistant Professor, Department of Political Science
University of Chicago

MICHAEL P. FISCHERKELLER
Doctoral candidate, Department of Political Science
The Ohio State University

RICHARD K. HERRMANN
Associate Professor, Department of Political Science
Director, Program in Foreign Policy Analysis, Mershon Center
The Ohio State University

ROBERT JERVIS
Adlai E. Stevenson Professor of International Relations,
Department of Political Science
Columbia University

YUEN FOONG KHONG
Fellow, Nuffield College
John G. Winant University Lecturer in American Foreign Policy
Oxford University

Sponsored by the Committee on International Peace and Security
of the Social Science Research Council

COUNTERFACTUAL THOUGHT EXPERIMENTS IN WORLD POLITICS

LOGICAL, METHODOLOGICAL, AND PSYCHOLOGICAL PERSPECTIVES

Edited by Philip E. Tetlock and Aaron Belkin

PRINCETON UNIVERSITY PRESS PRINCETON, NEW JERSEY

1

Counterfactual Thought Experiments in World Politics

LOGICAL, METHODOLOGICAL, AND PSYCHOLOGICAL PERSPECTIVES

PHILIP E. TETLOCK AND AARON BELKIN

THERE IS nothing new about counterfactual inference. Historians have been doing it for at least two thousand years. Counterfactuals fueled the grief of Tacitus when he pondered what would have happened if Germanicus had lived to become Emperor: "Had he been the sole arbiter of events, had he held the powers and title of King, he would have outstripped Alexander in military fame as far as he surpassed him in gentleness, in self-command and in other noble qualities" (quoted in Gould 1969). Social scientists—from Max Weber (1949) to Robert Fogel (1964)—have also long been aware of the pivotal role that counterfactuals play in scholarship on such diverse topics as the causes of economic growth and the diffusion of religious and philosophical ideas. Nevertheless, some contemporary historians still sternly warn us to avoid "what-might-have-been" questions. They tell us that history is tough enough as it is—as it *actually* is—without worrying about how things might have worked out differently in this or that scenario. Why make a difficult problem impossible? In this view (Fisher 1970; A. J. P. Taylor 1954), we do scholarship a grave disservice by publishing a volume on counterfactual reasoning. We are luring our colleagues "down the methodological rathole" in pursuit of unanswerable metaphysical questions that revolve around the age-old riddles of determinism, fate, and free will (Fisher 1970, 18).

The ferocity of the skeptics is a bit unnerving. Moreover, they are right that counterfactual inference is dauntingly difficult. But they are wrong that we can avoid counterfactual reasoning at acceptable cost. And they are wrong that all counterfactuals are equally "absurd" because they are equally hypothetical (Fisher 1970, 19). We can avoid counterfactuals only if we eschew all causal inference and limit ourselves to strictly noncausal narratives of what actually happened (no smuggling in causal claims under the guise of verbs such as "influenced," "responded," "triggered," "precipitated," and the like). Putting to the side whether any coherent and compelling narrative can be "noncausal," this prohibition would prevent us from drawing the sorts of "lessons from history" that scholars and policy makers regularly draw on such topical topics as the best ways to encourage economic growth, to preserve peace, and to cultivate democracy. Without counterfactual reasoning, how could we know whether state intervention accelerated growth in country *x*, whether deterrence prevented an attack on country *y*, or whether the courage of a young king saved country *z* from sliding back into dictatorship? Counterfactual reasoning is a prerequisite for any form of learning from history (cf. Tetlock 1991). To paraphrase Robert Fogel's (1964) reply to the critics of "counterfactualizing" in the 1960s, everyone does it and the alternative to an open counterfactual model is a concealed one.

This volume surveys the many roles that counterfactual arguments play in the study of world politics. A useful place to begin is by clarifying what we mean by counterfactual reasoning. A reasonably precise philosophical definition is that counterfactuals are subjunctive conditionals in which the antecedent is known or supposed for purposes of argument to be false (Skyrms 1991). As such, an enormous array of politically consequential arguments qualify as counterfactual. Consider the following rather representative sample of counterfactuals that have loomed large in recent scholarly and policy debates:

> If Stalin had been ousted as general party secretary of the communist party of the Soviet Union, the Soviet Union would have moved toward a kinder, gentler form of communism fifty-five years before it actually did;
>
> If Yeltsin had followed Sachsian fiscal and monetary advice in early 1992, Russian inflation in 1993 would have been a small fraction of what it was;
>
> If the United States had not dropped atomic bombs on two Japanese cities in August 1945, the Japanese would still have surrendered roughly when they did;
>
> If all states in the twentieth century had been democracies, there would have been fewer wars;
>
> If Bosnians had been bottlenosed dolphins, the West never would have allowed the slaughter of innocents in the Yugoslav civil war to have gone on so long.

The contributors to this volume approach counterfactual inference from both normative/epistemological and descriptive/cognitive science perspectives. The normative issues—which we explore in the next two sections of this chapter—focus on how students of world politics *should* use and judge counterfactual arguments. We break these issues into two categories:

(1) In what ways do counterfactual arguments advance our causal understanding of political events? Are such arguments—as the skeptics insist—merely forms of rhetorical posturing? Or can such arguments sensitize us to historical and theoretical possibilities that we might otherwise have ignored? Although we do not doubt that true believers often use counterfactuals to justify predetermined conclusions, it is a mistake to dismiss all such arguments as thinly veiled tautologies. We advance a provisional taxonomy of five constructive functions of counterfactual arguments in world politics, illustrating each with examples drawn from chapters in this volume.

(2) Once we settle on the appropriate purposes of counterfactual inference, what criteria should we use to distinguish plausible from implausible, insightful from vacuous arguments? Although we recognize the need for somewhat different criteria for distinctive "ideal-type" functions of counterfactuals, we see an even more pressing need to be explicit about the standards that scholars use in evaluating competing claims. There is an unfortunate tendency in the scholarly literature to oscillate between the extremes of dismissing dissonant counterfactuals as hopelessly speculative and of proclaiming favorite counterfactuals as self-evidently true, indeed as factual. This reaction is understandable, but unhelpful. The choice is typically not dichotomous; as we shall see, counterfactuals vary along a plausibility (or, if you are a Bayesian, subjective probability) continuum. If debates over competing counterfactuals are not to reduce to expressions of theoretical or ideological taste, we need to articulate standards of evidence and proof that transcend rival schools of thought. In this spirit, we advance a provisional list of six standards for judging counterfactual claims, illustrating each standard with examples drawn from later chapters.

The final section of this chapter shifts the focus from "how *should* we generate, use, and judge counterfactual arguments?" to "how *do* we generate, use, and judge counterfactual arguments?" One key cognitive-science question concerns when people are prone to think about possible worlds. Of the infinity of past events that people could "mentally undo" and insert as antecedents into counterfactual arguments, why do they devote so much attention to certain causal candidates and so little to others (Kahneman and Miller 1986; Commentary 2, Olson, Roese, and Deibert)? A natural next question concerns when people are likely to be persuaded by counterfactual claims concerning the consequences of altering particular antecedents. Given that people have no way of directly determining what would have happened in these hypothetical worlds, why do they defer to some counterfactual arguments but disdain others (Commentary 1, Turner)? Finally, we explore the potential for double standards in so subjective a domain as thought experiments. Is there evidence of cognitive and motivational biases in how people judge claims about possible worlds, tendencies to raise standards of evidence and proof for dissonant counterfactuals but to lower standards for claims consonant with one's beliefs and goals?

Normative Issues in Evaluating Counterfactual Claims

Our contributors generally agree that counterfactual reasoning is unavoidable in any field in which researchers want to draw cause-effect conclusions but cannot perform controlled experiments in which they randomly assign "subjects" to treatment conditions that differ only in the presence or absence of the hypothesized cause. Try though we do to control statistically for confounding variables in large-*N* multivariate studies or to find matching cases in comparative designs or to search for the signature of hypothesized causes in process-tracing studies, the potential causes are simply too numerous and too interrelated in world politics to permit complete escape from counterfactual inference. Researchers must ultimately justify claims that a given cause produced a given effect by invoking counterfactual arguments about what would have happened in some hypothetical world in which the postulated cause took on some value different from the one it assumed in the actual world (Fogel 1964; Fearon 1991).

The consensus among our contributors, however, begins to unravel beyond this point. They emphasize distinctive, albeit largely complementary, functions of counterfactual reasoning. The arguments they present have persuaded us to adopt a stance of epistemic pluralism that acknowledges the variety of ways in which counterfactual arguments can prove enlightening and the need for different standards in judging counterfactuals that serve different scholarly goals. We organize these distinct styles of counterfactual argumentation into five ideal types:

1. *Idiographic case-study counterfactuals* that highlight points of indeterminacy at particular junctures in history (reminding us of how things could easily have worked out differently and of how difficult it is to apply abstract hypothetico-deductive laws to concrete cases);
2. *Nomothetic counterfactuals* that apply well-defined theoretical or empirical generalizations to well-defined antecedent conditions (reminding us that deterministic laws may have been at work that were invisible to the original historical actors as well as to contemporary scholars who insist on a radically idiographic focus on the particular);
3. *Joint idiographic-nomothetic counterfactuals* that combine the historian's interest in what was possible in particular cases with the theorist's interest in identifying lawful regularities across cases, thereby producing theory-informed history;
4. *Computer-simulation counterfactuals* that reveal hitherto latent logical contradictions and gaps in formal theoretical arguments by rerunning "history" in artificial worlds that "capture" key functional properties of the actual world;
5. *Mental-simulation counterfactuals* that reveal hitherto latent psychological contradictions and gaps in belief systems by encouraging people to imagine possible worlds in which causes they supposed irrelevant seem to make a difference, or possible worlds in which causes they supposed consequential seem to be irrelevant.

Five Styles of Counterfactual Argumentation

1. Idiographic

Several authors use counterfactuals to explore "possibility-hood"—whether history had to unfold as it did. For instance, Breslauer (Chapter 3) explores the several junctures in the history of the Soviet Union that have sparked the most intense counterfactual debate within the expert community: Was the Bolshevik revolution inevitable given the Russian defeat in World War I? Was Stalinism inevitable given the vanguard-party legacy of Leninism? Was Gorbachevism inevitable given the repressive stagnation of Brezhnevism? And was the disintegration of the Soviet Union inevitable given the liberal reforms of Gorbachevism? Khong (Chapter 4) attempts to assess whether any conceivable British prime minister would have adopted a policy of appeasement toward Nazi Germany, at least up to March 1939. Herrmann and Fischerkeller (Chapter 6) examine several counterfactual controversies in which the positions taken by policy makers on "what would have happened?" shaped American policy toward Iran during the Cold War. Lebow and Stein (Chapter 5) construct an exhaustive inventory of the counterfactual beliefs that apparently guided American and Soviet policy during the Cuban missile crisis—the crisis during which, it is often asserted, the world "came closer" than ever before or since to nuclear war.

These diverse applications all use counterfactuals to focus on "conceivable" causes that could have easily redirected the path-dependent logic of events (cf. Hawthorn 1991; Chapter 2, Fearon). In each case, the investigators want to know what was historically possible or impossible within a circumscribed period of time and set of relations among political entities. To make this determination, they draw upon combinations of: (a) in-depth case-specific knowledge of the key players, their beliefs and motives, and the political-economic constraints under which they worked; and (b) general knowledge (nomothetic propositions) concerning cause-effect relations in human behavior and political-economic systems. Moreover, our case-study authors seem to agree that counterfactual speculation should be constrained by some form of "minimal-rewrite-of-history" rule that instructs us to avoid counterfactuals that require "undoing" many events—counterfactuals that, for instance, ask us to imagine a democratic Soviet Union at the end of World War II or Soviet possession of strategic nuclear superiority at the time of the Cuban missile crisis. A more fruitful way to proceed is to ask what could have worked out differently if we introduce *easily imagined* variations into the causal matrix of history. Might the murderous tyranny of Stalin have been averted if Trotsky had not gone duck hunting, caught a cold, and missed a key politburo meeting or if Bukharin had been a savvier politician? Might World War II have been nipped in the bud if British opponents of appeasement had had one or two additional cabinet seats during the Munich crisis? And might World War III have been triggered in October 1962 if Kennedy had followed the advice of his more hawkish advisors and immediately ordered air strikes against Soviet missile sites in Cuba?

These idiographic counterfactuals are not idle exercises in social-science fiction; they are a useful corrective to simple deterministic forms of theory. They compel us either to abandon determinism by acknowledging the role of chance or to abandon simplicity by acknowledging that factors outside the purview of our deterministic models—viruses, skillful or inept leadership, group dynamics, a well-timed or ill-timed persuasive argument—can decisively alter the course of events.

Beyond their heuristic contribution to social science theory, idiographic counterfactuals are an integral part of the process of passing moral judgment on individual leaders and even entire political systems such as Marxism-Leninism. We rely on them in attributing responsibility (Hart 1961). Would a reasonable person, confronted by these circumstances, have acted differently? Should a particular leader be praised for performance above the norm (spectacular prescience or courage) or condemned for performance below the norm (stubborn refusal to recognize trends apparent to others or cowardly failure to protest immoral conduct)? Neville Chamberlain, John Kennedy, Nikita Khrushchev, and Lyndon Johnson are all, in a sense, in the docket with their reputations as wise leaders hanging in the balance on counterfactual judgments of what they could or should have done at certain junctures in history.

2. *Nomothetic Theory-Testing*

Whereas idiographic investigators are interested in conceivable causes that they can readily imagine taking on different values within a specific historical context, nomothetic investigators usually show little or no concern for the plausibility of switching the hypothesized counterfactual antecedent on or off in any given context. From this perspective, counterfactuals are the inevitable logical by-products of applying the hypothetico-deductive method to an historical (nonexperimental) discipline such as world politics. Whenever we combine a well-defined Hempelian covering law (say, relating money supply to inflation) with well-defined antecedent conditions (the Russian economy in January 1992), we can deduce specific counterfactual conclusions (e.g., if the Russian central bank had adopted this or that monetary policy, then, *ceteris paribus*, inflation would have taken on this or that value). Note that these counterfactuals are in no way constrained by the historical plausibility of the Russian central bank adopting one or another policy. The counterfactual "predictions" follow from the context-free logic of macroeconomic theory, not from the context-bounded logic of what was psychologically or politically possible at that juncture in Russian history. Adopting Fearon's (Chapter 2) terminology, these nomothetic counterfactuals invoke miracle causes. Even if our theory requires us to posit an extremely implausible hypothetical world, we do what our theory tells us to do. The goal is not historical understanding; rather, it is to pursue the logical implications of a theoretical framework. For instance, Kiser and Levi (Chapter 8) note that influential sociological theories of revolution imply that if there had been a large, educated middle class in the France of 1789 or in the Russia of 1917, revolution would not have occurred. Russett's (Chapter 7) democratic-peace hypothesis implies that if all states in the twentieth century had been democracies, war would have been less frequent. Keohane's (1984) theoretical work on regimes claims that if international regimes did not exist, there would be markedly less international cooperation. Waltz's (1979) structural neorealism implies that if we transformed a multipolar state system (e.g., pre–World War I Europe) into a bipolar one, the stability of the system would have increased.

What makes these counterfactuals anything more than dogmatic reassertions of faith in a theory that stipulates "cause *x* facilitates outcome *y* and, in the absence of cause *x* and all other things being equal, the likelihood of *y* diminishes by some amount"? A fuller answer to this question emerges in our later discussion of the statistical and projectability tests of counterfactuals. For now, it must suffice to note the root difficulty: namely, history is a terrible teacher. Key events occur only once, whereas for purposes of valid causal inference we would like to rerun history many times and to examine the resulting distribution of outcomes in contingency tables that reveal how strongly causes and effects covary. But time-machine experimentation of this sort is impossible, so we are stuck with the covariation data available in the real world (a world in which the numbers of democratic states and wars are both constants). We then have to rely on the imperfect statistical means at our disposal to estimate the degree to which democracy inhibits war, controlling as best we can for confounding variables. From the Dawesian perspective (Commentary 3), the democratic-peace counterfactual can be only as true as the covariation data in the real-world contingency tables permit. In King, Keohane, and Verba's (1994) framework, there is additional latitude for learning about the truth-status of counterfactuals (see also Chapter 7, Russett). As good theorists, it is incumbent upon us to go beyond mere observations of covariation and to stipulate the causal mechanisms underlying the democratic peace and to derive a host of testable predictions from these hypothesized mechanisms. For instance, if heightened accountability constraints on leaders are responsible for the democratic peace, what independent evidence do we have of the workings of this hypothesized cause? Are democratic leaders who advocate war against fellow democracies more likely to fail than their less bellicose colleagues? Do we see more references to accountability constraints in the private deliberations of democratic than nondemocratic leaders? The more elaborate the network of corroborative correlational evidence, including time-lagged and partial correlations, the greater our justifiable confidence in the nomothetic counterfactual.

3. Idiographic-Nomothetic Synthesis

The tension between idiographic disciplines (history and area studies) and nomothetic disciplines (general social science) is well known and need not be belabored. A not uncommon way of proceeding is to acknowledge that the idiographic and nomothetic represent complementary "ways of knowing" that may in the fullness of time be conceptually integrated, but do not hold your breath. It is worth noting, however, that such conceptual integration is the norm in natural history, where there is much less controversy than in the social sciences over what counts as a well-established statistical or theoretical generalization.

Our favorite example of idiographic-nomothetic symbiosis is the manner in which biological and physical scientists have gone about deriving and testing rival hypotheses concerning the extinction of dinosaurs. Perhaps the most influential hypothesis is the doomsday-asteroid conjecture which, in counterfactual form, runs as follows: "If a six- to twelve-mile-wide asteroid had struck the Earth at a velocity of approximately 44,000 miles per hour sixty-five million years ago, then a host of predictions would follow (including the size of the crater, the effects on the atmosphere and climate, the distribution of various trace elements in particular geological strata, antipodal volcanism, . . .)." This line of work captures the best in both the idiographic and nomothetic traditions. Investigators focus on a well-defined "conceivable" cause (meteors and asteroids hit our planet frequently over long stretches of time) but rely heavily upon deductive theory, empirical observations, and computer simulations to assess the soundness of the connecting principles that permit us to deduce empirical consequences such as climate change of sufficient magnitude to wipe out the dinosaurs. Investigators also try to tease apart testable predictions from rival hypotheses such as "endogenous volcanism alone is sufficient to account not only for this specific mass extinction but for nine of the ten other mass extinctions in the

fossil record over two billion years." As a result of this vigorous research program, many scientists argue that a once highly speculative counterfactual conjecture is now better viewed as a quite-probable fact of natural history—yet another illustration of how blurry the boundary between factual and counterfactual can be (Chapter 6, Herrmann and Fisherkeller).

There are no idiographic-nomothetic syntheses of comparable scope and sweep in world politics. But there are some elegant demonstrations of how one can weave together idiographic and nomothetic objectives—in particular, by the game theorists in this volume. Bueno de Mesquita and Weingast both use game-theoretic models to enhance our understanding of particular historical episodes (Philip Augustus versus the Pope; medieval merchants versus towns; federal bureaucrats versus Congress), to identify intriguing cross-case regularities, and to make predictions about how behavior will change as a lawful function of alterations in the probabilities or payoffs attached to courses of action. In so doing, the game theorists remind us that social scientists are not the only creatures roaming this planet capable of thinking counterfactually. Policy makers do it all the time, constructing mental representations of how others would respond to one or another move and making decisions on the basis of those mental models. Policy makers can identify equilibrium solutions (solutions in which no one stands to gain from unilateral defection) only by computing off-the-path behavioral (OTPB) expectations concerning what would happen if they or the other side acted differently in response to a given move. Assuming both sides act rationally and stay on the equilibrium path, these OTPB expectations eventually become counterfactual assertions about what would have happened under this or that contingency (Chapter 9, Bueno de Mesquita; Chapter 10, Weingast). The mental representations of these now-counterfactual worlds were once, however, causally consequential; they constrained rational decision makers to go down particular branches of the game-tree.

Game theorists integrate the idiographic and nomothetic by applying "strong theory"—expected utility maximization and criteria for identifying equilibrium strategies—to complex historical situations that can then be understood by modeling the options available to each side and the expected payoffs associated with all logically possible combinations of moves. In judging what else could plausibly have happened, game theorists use nomothetic laws to answer the idiographic question: How much history do I have to rewrite to "undo" a particular policy? If the counterfactual simply shifts us from one equilibrium path to another (as is possible in games with multiple equilibria), the counterfactual does no violence to the rational-actor axioms of the underlying theory and may be quite acceptable. But if the counterfactual requires us to imagine a world in which, for stochastic reasons ("trembling hand") or psychological reasons ("bounded rationality," motivational perversity), players stray from an equilibrium to a nonequilibrium path so that one or both are worse off than they otherwise could be, the counterfactual is suspect. These ground rules for judging the permissibility of possible worlds are commendably precise, albeit rather procrustean. There is no guarantee that history is efficient in the sense of quickly identifying equilibrium solutions; history may be better viewed as a "path-dependent meander" (March and Olson 1995) in which accidents, fortuitous opportunities, and miscalculations often lead us into culs-de-sac from which it is difficult, even impossible, to extricate ourselves.

4. Pure Thought Experiments: Logical Proofs and Computer Simulations

Our contributors often use counterfactuals to reinforce a causal argument (be it an idiographic one concerning the impact of a particular belief, person, or policy, or a nomothetic one concerning causal processes that theoretically transcend context). But they also sometimes use counterfactuals to reveal previously hidden contradictions or ambiguities in the logical structure of the causal arguments that others have advanced.

Using counterfactuals to probe the logical completeness and internal coherence of claims is commonplace in mathematics, the physical sciences, and economics. A prototype is Euclid's elegant proof that the number of prime numbers must be infinite because if we take the counterproposal seriously, we are compelled to make contradictory claims. For example, if and only if the number of prime numbers were finite, then there would exist a nonprime number x such that x equals the product of all primes plus 1 ($x = (p_1 p_2 \ldots p_n) + 1$). But if this were true, x as a nonprime number must by definition be directly factorable into either nonprimes or primes and, if factorable into nonprimes, those nonprimes must eventually be factorable into primes. But this is impossible given the method of constructing x, so the number of primes must be infinite and the antecedent must be false.

We know of no comparable *reductio ad absurdum* in world politics or indeed of thought experiments that are as decisive in shaking theoretical convictions as those of Galileo and Einstein in physical science or of Ricardo, Coase, and Arrow in economic theory. But we do see some interesting parallels with the computer simulations of complex adaptive systems that Cederman and Fearon discuss in their respective chapters. One interpretation of these simulations is that they highlight logical lacunae in currently influential approaches to world politics. The qualification "one interpretation" is critical; one is not obliged to accept this interpretation for the simple reason that the simulation-based counterfactuals lack the "if and only if" delivering power of rigorous mathematical proofs in well-defined axiomatic systems. For example, one could argue that if balancing were inevitable in anarchic international systems, then global hegemons would not emerge in simulated worlds which, according to Cederman (Chapter 11), capture the key functional attributes of anarchy within a neorealist framework. But because hegemons do emerge, and emerge especially frequently when defense-dominance prevails (an additional unwelcome surprise for some theorists), this neorealist prediction may (not must) be wrong. Cederman's simulations of artificial histories suggest that we may have just been lucky that an Alexander or Hitler or Napoleon has not yet conquered the world! Or, shifting to Fearon's chapter, one could argue that if long-term forecasting were possible in complex interdependent systems, then we could predict the long-term consequences of minor variations in initial settings for cellular automata. But because we cannot make accurate long-term predictions even in these simple, well-understood systems, perhaps long-term predictability also breaks down in the much more complex and poorly understood domain of world politics. These simulation-driven counterfactuals are not deductively decisive but they are intellectually seductive. They nudge us gently toward the conclusion that something is awry with key assumptions that serve as starting points for influential analyses of security issues.

5. Mental Simulations of Counterfactual Worlds

Not all counterfactual simulations of possible worlds need run through the logical structures of computer programs; some run through the psychological structures of the human mind. The classic thought experiments of physicists and economists illustrate the point in the abstract, but it is possible to make the same point with examples more directly relevant to world politics. Asking people to imagine and work through the detailed implications of hypothetical worlds is a powerful educational and rhetorical tool. Like their formal epistemological kin (logical proofs and computer simulations), mental simulations can highlight critical contradictions and ambiguities in one's own and others intellectual positions (see also Turner's notion of spotlight counterfactuals in Commentary 1). As Kahneman (1995) points out, mental simulations derive their persuasive force and power to surprise by revealing previously unnoticed tensions between explicit, conscious beliefs and implicit, unconscious ones. In this sense, people discover aspects of themselves in mental simulations that would otherwise have gone undiscovered. We find it useful to distinguish three specific ways in which mental simulations can yield insights into our own thought processes: by revealing double standards in moral judgment, contradictory causal beliefs, and the influence of unwanted biases such as certainty of hindsight.

COUNTERFACTUAL MORALITY TALES

Mental simulations can compel people to acknowledge embarrassing or even shameful inconsistencies in their application of moral rules. The paradigmatic example is the identity-substitution thought experiment that manipulates either the perpetrator or victim of a deed and asks the audience to contemplate whether they had the same emotional reaction to what actually happened as they would have had to various hypothetical events. For instance: "If Bosnians were bottlenosed dolphins [Rwandans white, Chechnyans Lithuanians . . .], we never would have tolerated the slaughter of innocents so long." Insofar as the audience detects a discrepancy in their reactions to the two scenarios, and insofar as the audience firmly believes that the mentally manipulated cause should be irrelevant, the audience will deem the discovery of a differential emotional reaction to be a disturbing *fact* about themselves. Moreover, the thought experiment is easily translated into an actual experiment. For example, survey researchers often perform actual identity-substitution experiments to gauge the influence of "socially undesirable" causes of policy preferences that, it is assumed, people would not be willing to acknowledge if they were asked directly (Sniderman, Brody, and Tetlock 1991).

COUNTERFACTUAL CONSISTENCY PROBES

Here the mental simulation reveals contradictions between causal beliefs that may have previously coexisted peacefully within a belief system. The paradigmatic example is the syllogism that traces through the logical implications of one set of beliefs to the point where the contradiction becomes undeniable. For example:

> If you really believe that there is so much indeterminacy at the micro level (battles, firms, bureaucratic subunits of government), you cannot plausibly argue for so strongly deterministic a position at the aggregate or macro level (wars, economies, government decisions);
>
> If you commit yourself to an extreme structural-realist position that denies any significant causal role to domestic politics in shaping properties of the international system, then you would have to believe that even if the Soviet Union had been a democracy in 1945, the Cold War would still have occurred.

The thinker then has the dissonance-reduction options of changing one or both sets of beliefs, introducing new cognitions that neutralize the contradiction, or disengaging from the simulation exercise by simply ignoring the contradiction (cf. Abelson 1959; Festinger 1957).

It is important to be explicit about the likely long-term impact of repeated counterfactual-consistency probing of expert belief systems. Certain belief systems are much more vulnerable to this type of conceptual challenge: specifically, belief systems organized around strongly deterministic claims (if *x*, then *y* must occur) and strongly exclusionary claims (*x* and only *x* influences *y*). The greater vulnerability can be traced to the greater ease of generating and justifying "could" versus "would" counterfactuals. A "could" counterfactual merely requires showing that there is at least one plausible story about a possible world in which *x* did not automatically lead to *y* or in which some other cause, *z*, also influenced *y*. By contrast, an unqualified "would" counterfactual requires showing that an outcome would have occurred in all possible worlds that pass some threshold of plausibility. The burden of proof is obviously much higher in the latter case.

COUNTERFACTUAL EXERCISES AS DEBIASING TOOLS AND MEANS OF STIMULATING THE IMAGINATION

Retrospective scenario generation is mental simulation in which the goal is to prevent premature cognitive closure induced by certainty of hindsight. Some scholars (e.g., Weber 1949) have long suspected, and cognitive psychologists have recently demonstrated (e.g., Fischhoff 1975; Hawkins and Hastie 1990), that "outcome knowledge" contaminates our understanding of the past. Once people learn the outcome of an event, they not only perceive that outcome as more likely ex post than they did ex ante (which might be defended as a rational Bayesian updating of subjective probabilities), they often fail to remember their ex ante assessment of what was and was not likely to happen. Backward and forward reasoning in time are, in this cognitive sense, deeply asymmetrical.

Counterfactual thought exercises can check the "creeping determinism" of certainty of hindsight. Asking people to think of how things could have worked out differently becomes a means of preventing the world that did occur from blocking our views of the worlds that might well have occurred if some antecedent condition had taken on a different value. Indeed, there is a small "debiasing" literature in experimental psychology that assesses the usefulness of encouraging people to imagine that the opposite outcome occurred (Fiske and Taylor 1991). There is also a small literature in literary studies on the concepts of sideshadowing, foreshadowing, and backshadowing in narratives that makes a strikingly similar normative point (M. A. Bernstein 1994; Morson 1994). Sideshadowing calls attention to what could have happened, thereby locating what did happen in the context of a range of possibilities that might, with equal or even greater likelihood, have taken place instead. Sideshadowing serves as a valuable check on foreshadowing (the tendency, in extreme form, to reduce all past events to harbingers of the future) and backshadowing (the even more insidious tendency to judge historical actors as though they too should have known what was to come).

Bernstein cautions us about the dangers of adopting a condescending "backshadowing" attitude toward participants in past events—such as the victims of the Holocaust—that were neither inevitable nor perhaps even predictable.

Of our contributors, Weber (Chapter 12) is most concerned with the potentially liberating effects of allowing our counterfactual imaginations freer rein than they are usually given. He suggests that a partial explanation of why international relations theorists are so often surprised by events is a failure of divergent thinking—a failure to give due weight to the variety of possible pasts that could have occurred as well as to the variety of possible futures that might yet occur. Confronted by a complex probabilistic world in which the tape of history only runs once, prudent decision makers should entertain multiple plausible scenarios of how events might have unfolded and might yet unfold, hedging their policy bets accordingly (Schoemaker 1991). Weber also criticizes methodologists and epistemologists, including us, who try to constrain our counterfactual imaginations by invoking the sorts of plausibility tests advanced in the next section. We agree with Weber that deterministic tunnel vision is a serious problem but worry about the viability of the proposed solution. Weber may be right that lack of creative divergent thinking is a more serious deficiency in world politics than proliferation of false hypotheses, but if the field took Weber's advice to heart, the opposite might well soon be the case. Any self-respecting academic community must offer the attentive public criteria for distinguishing scenario snake oil from serious scholarship.

Six Criteria for Judging Counterfactual Arguments

There should now be no doubt that scholars use counterfactual arguments for a variety of distinct, albeit interrelated, purposes. It should also come as no surprise that there is no single answer to the question of what counts as a good counterfactual argument. The obvious rejoinder is, "Good for what?" A counterfactual that is idiographically incisive (advances our understanding of a particular case) might be nomothetically banal (devoid of interesting theoretical implications) and vice versa. A counterfactual grounded in an elegant computer simulation might blow a gaping logical hole in an influential theoretical argument but tell us precious little about the actual world it supposedly simulates. A counterfactual that stimulates us to think of new hypotheses might run afoul of the received wisdom on what counts as a trivial or influential cause.

Given the diverse goals that people have in mind when they advance counterfactual arguments—from hypothesis generation to hypothesis testing, from historical understanding to theory extension—our contributors convinced us that the quest for a one-size-fits-all epistemology is quixotic. Different investigators will inevitably emphasize somewhat different criteria in judging the legitimacy, plausibility, and insightfulness of specific counterfactuals. It would be a big mistake, however, to confuse epistemic pluralism (which we accept up to a point) with an anything-goes subjectivism (which we reject and which would treat all counterfactual claims as equally valid in their own way). The study of world politics has suffered from a lack of self-consciousness about the counterfactual underpinnings of causal claims (Fearon 1991). Indeed, different schools of thought sometimes seem precariously close to establishing their own implicit norms for deciding what is and is not a "trivial" argument (Chapter 12, Weber)—an outcome that would be disastrous because it would permit rival schools to disengage altogether from constructive arguments with each other. A science or, more modestly, quasi science of world politics is possible if and only if advocates of conflicting hypotheses embrace at least some common standards for judging the plausibility of each other's counterfactual claims. Otherwise, we are fated to talk past each other.

To avoid this fate, we advance six normative criteria for judging counterfactual arguments that appear to command substantial cross-disciplinary support. To be sure, we do not expect universal consent; we do seek, however, to initiate a sustained conversation within the research community on what should count as a compelling counterfactual argument—a conversation that will allow us to explore the strengths and weaknesses of specific standards in the abstract, in isolation from the dominant debates of the moment (when the temptation to play favorites is often irresistible). There should, moreover, be plenty to talk about. Each standard we propose is open to some interpretation. Certain standards will provoke resistance from those who denounce it as impossible (Chapter 3, Breslauer) or undesirable (Chapter 12, Weber) or irrelevant (Chapter 5, Lebow and Stein). And some standards will clash with each other. For example, consistency with well-established historical fact sometimes conflicts with consistency with well-established statistical or theoretical generalizations. There are at present no generally accepted principles for adjudicating such disputes and we do not claim to offer a well-defined "method of counterfactual argument" that researchers can deploy in an off-the-shelf fashion to solve any and all problems.

With these disclaimers, we list six attributes of the ideal counterfactual thought exercise (where "ideal" means most likely to contribute to the ultimate social-science goals of logically consistent, reasonably comprehensive and parsimonious, and rigorously testable explanations that integrate the idiographic and the nomothetic):[1]

[1] Readers may wonder how these six criteria for good counterfactual reasoning map onto the five ideal-type patterns of counterfactual reasoning sketched earlier. Although the answer is complex and the subject of some disagreement among contributors, we can offer the following observations:

(1) The first two criteria—logical clarity and cotenability—have widespread acceptance

1. *Clarity*: Specify and circumscribe the independent and dependent variables (the hypothesized antecedent and consequent);

2. *Logical consistency or cotenability*: Specify connecting principles that link the antecedent with the consequent and that are cotenable with each other and with the antecedent;

3. *Historical consistency (minimal-rewrite rule)*: Specify antecedents that require altering as few "well-established" historical facts as possible;

4. *Theoretical consistency*: Articulate connecting principles that are consistent with "well-established" theoretical generalizations relevant to the hypothesized antecedent-consequent link;

5. *Statistical consistency*: Articulate connecting principles that are consistent with "well-established" statistical generalizations relevant to the antecedent-consequent link;

6. *Projectability*: Tease out testable implications of the connecting principles and determine whether those hypotheses are consistent with additional real-world observations.[2]

across ideal types and are perhaps the best candidates for the status of universal minimum standards;

(2) The third criterion—consistency with well-established historical fact (also known as the minimal-rewrite rule)—is carefully observed by most idiographic researchers (ideal type 1) but frequently ignored by many nomothetic researchers (ideal type 2);

(3) The fourth, fifth, and sixth criteria—consistency with well-established theoretical and statistical generalizations and projectability—are more widely acknowledged among nomothetic than among idiographic researchers (although a significant contingent of idiographic researchers do apply these standards in their own work).

[2] Some logicians (Stalnaker 1984; Lewis 1973) have proposed a seventh test—the semantics of possible worlds—for judging the truth or falsehood of counterfactual claims. To test a proposition of the form "if p, then q," possible-worlds semantics directs us to do three things: (1) identify the set of possible worlds $\{p\}$ in which the counterfactual antecedent, p, is true; (2) identify that possible world $\{p^1\}$ that is "closest" to the actual world; (3) determine whether that possible world $\{p^1\}$ falls in the intersection of the set of possible worlds in which p is true, $\{p\}$, and the set of possible worlds in with q is true, $\{q\}$. We should judge "if p, then q" to be true if and only if the "closest" p falls in the intersection of $\{p\}$ and $\{q\}$. In other words, we should judge "if p, then q" to be true if and only if the closest world in which p is true is also a world in which q is true.

This logical calculus provides an elegant framework for evaluating counterfactual claims. It assumes, however, a vastly more sophisticated knowledge of the causal workings of the world than social scientists currently possess (or are likely to possess anytime in the next century). We need to partition the universe of possible worlds into overlapping sets, to locate the actual world in the universe of possible worlds, and to quantify the "distance" between the actual world and each possible world. Not surprisingly, none of our contributors could implement this test. We differ from Lebow and Stein (Chapter 5) in that we distinguish the Lewis-Stalnaker approach from the minimal-rewrite rule. Even if the antecedent p is an historically implausible miracle cause (Chapter 2, Fearon), the closest world in which p is true could still be a world in which q is true, in which case the counterfactual violates the minimal-rewrite rule but passes the Lewis-Stalnaker test.

1. Well-Specified Antecedents and Consequents

Our first recommendation might strike readers as a tad obvious. Like actual experiments, thought experiments should manipulate one cause at a time, thereby isolating pathways of influence. Although excellent advice in principle, implementing it is often deeply problematic. There is no way to hold "all other things equal" when we perform thought experiments on social systems that are densely interconnected (cf. Jervis 1993; Commentary 4, Jervis). To invoke the terminology of experimental design, we cannot manipulate the "independent" variable in interconnected systems without creating ripple effects that alter the values taken on by other potential causes in the historical matrix, thereby creating "confounding" variables that render interpretation of the original thought experiment problematic.

At this juncture, radical wholists take the systems-theory argument even further and insist that if we want to advance coherent and defensible counterfactuals, we will have to reconstruct an entirely new hypothetical world for each new counterfactual proposition—a new world that specifies all other things that would also have to change in order to accommodate the hypothesized antecedent (otherwise the counterfactual is underspecified). This position strikes us as too extreme. Causal interconnectedness is a matter of degree. In the words of one systems theorist: "Everything is connected but some things are more connected than others. The world is a large matrix of interactions in which most of the entries are very close to zero" (Pattee 1973, 23). The analytical challenge then becomes estimating interconnectedness and designing our counterfactual thought experiments with due consideration for the complexities created by interconnectedness. Sometimes we will discover that the wholists are right: the causal antecedent that we mentally manipulated is so deeply embedded in a recursive network of causation that simply positing that "if cause x took on a different value, then y" is deeply uninformative. Consider two examples:

(1) Cederman (Chapter 11) criticizes the structural-realist claim advanced by Mearsheimer (1990, 14) that "*ceteris paribus*, war is more likely in a multipolar system than a bipolar one." In Cederman's view, other things probably cannot be held equal in the post–World War II case, and Mearsheimer glosses over the problem by failing to articulate what else would have had to be different in counterfactual post–World War II systems in which multipolarity prevailed. As soon as we try to specify the alternatives to a bipolar world more precisely, we begin to appreciate the need for domestic-political boundary conditions on the polarity counterfactual. For instance, the identity of the third power—be it Great Britain, France, or China—might matter. Given the special historical relationship between Lon-

don and Washington, it is not intuitively obvious that a tripolar world composed of the Soviet Union, the United States, and Britain would have been less stable than the actual bipolar world.

(2) Shifting to economic history, Gould (1969) complains about the counterfactual, "If the Industrial Revolution had not occurred, the British standard of living would have been lower than it was." He observes:

> We cannot decide what we must subtract from the real past along with the Industrial Revolution. . . . In order to know what would have happened to income per head had the Industrial Revolution not occurred we need to know, amongst other things, what in such circumstances, would have happened to population. But to know what, in those same circumstances, would have happened to population we need to know, amongst other things, what would have happened to income per head.

It is not clear that we can escape this "vicious circle." At a minimum, we need to clarify the counterfactual antecedent by creating a "compound" (e.g., if the Industrial Revolution had not occurred *and* if British population grew at the same rate between 1750 and 1850, then . . .) and by specifying what else would have to be different about this hypothetical Britain from which we have now "subtracted" two fundamental causal processes: industrial growth and rising population.

These arguments are grist for the wholists mill. But in other cases, causal interconnectedness seems much thinner. Perhaps one reason why assassinations attract so much counterfactual attention is that it is so easy to imagine "getting away with" changing only a few causal antecedents and producing a consequential result. It requires little rewriting of history to posit hypothetical worlds in which Oswald missed his target or in which Kennedy chose to ride through Dallas in a car with a bulletproof roof. These possible worlds are only a muscle twitch or nightmare removed from the actual world.

We run into similar conceptual problems on the consequent side of counterfactuals. Consider a variation of Pascal's conjecture on the causal impact of Cleopatra's nose on the course of Western history. Fearon (Chapter 2) concedes that "if Cleopatra had an unattractively large nose, World War I might not have occurred" but argues that the counterfactual hardly belongs in any reasonable explanatory account of World War I. If Cleopatra's nose were that consequential, the hypothetical world of 1914 is almost certainly not just a minor variant of the actual world of 1914, but rather a radically different world in which the nonoccurrence of World War I is but one of countless points of difference that go back 2,000 years. There might also be no Germany or Great Britain. Fearon proposes, as a pragmatic rule of evidence, that we seriously consider only those counterfactuals in which the antecedent seems likely to affect the specified consequent *and* very little else. This argument invokes a surgical-strike model of counterfactual inference in which we not only manipulate one thing at a time, we give priority attention only to those causes specifically relevant to the consequent of interest (if and only if the hypothesized causal variable takes on the value x', then the effect occurs *and* everything else in the hypothesized world is pretty much identical to the actual world).

Fearon's proposal is open to challenge on the ground that it arbitrarily rules out causes that, because of their location in complex systemic networks of causation, do not have effects that can be conceptually isolated. For our part, we see no easy resolution of the tension between the desire of methodologists to "hold other things equal" and the insistence of latter-day Leibnizians that once we tamper with one element from the past, we have to trace through the causal implications for all other elements, in effect creating a full-fledged alternative world for each counterfactual. The argument is best engaged on a case-by-case basis, with a minimum of metaphysical posturing. Investigators should obviously be sensitive to systemic effects and be precise about the implications of implementing their hypothesized causes in hypothetical worlds. In some cases, the grounds for suspecting systemic effects will be weak and the counterfactual exercise can approximate the austere parsimony of the thought experiment; in other cases, the grounds for suspecting systemic effects will be powerful and counterfactual exercises will acquire the rich narrative trappings of scenario generation, with detailed stories and subplots elaborated around why certain historical paths were not taken and what would have had to be different to activate them (Chapter 12, Weber).

2. Cotenability: Logical Consistency of Connecting Principles

Every counterfactual is a condensed or incomplete argument that requires connecting principles that can sustain, but not imply, the conditional claim (Goodman 1983). When explicitly articulated, these connecting principles are often complex, even in the case of such seemingly simple counterfactuals as "If the match had been scratched, it would have lighted." The connecting principles specify, within reasonable limits, everything else that would have to be true to sustain the counterfactual, including the necessary amount of friction generated by the scratch, the chemical composition of the match, the absence of water, the presence of oxygen, and so forth.

In our view, connecting principles should satisfy three minimal criteria. They should be specified reasonably precisely, be consistent with each other, and be consistent with both the antecedent and consequent. Unfortunately, as several contributors point out, counterfactual arguments in world politics often fail the first test so badly that it is impossible to tell how well they might have fared against the second and third tests. Focusing on the Cuban missile crisis, Lebow and Stein (Chapter 5) note that liberal "revisionists . . . provide no compelling justification for their expectation that had Kennedy made a secret overture to Khrushchev before choosing the blockade, Khrushchev would have responded positively." It is just as plausible that he would have stood firm and accelerated the construction of the missile sites—as the Soviet military in Cuba did initially in response to the blockade. Although some liberal revisionists do advance the connecting-principle rationale that it would have been easier for Khrushchev to back down in the absence of a public confrontation, they cannot rebut the counterargument that Khrushchev needed a serious confrontation to justify a withdrawal to hard-liners in the politburo. They cannot do so because they lack a sound basis for specifying when the Soviet political leadership would have responded in an accommodative or confrontational fashion.

Conservative revisionists have similar problems. They are fond of counterfactuals in which a president who displayed greater resolve prevented the missile crisis of 1962. In this view, Khrushchev doubted Kennedy's resolve for two reasons: the president's poor performance, and Khrushchev's view of Americans as "too soft, too liberal, and too rich to fight." The counterfactual hypothesizes that Khrushchev would not have sent missiles to Cuba if Kennedy had displayed greater resolve at the Bay of Pigs, at the Vienna summit, and in Berlin. It does not specify, however, how presidential displays of resolve would have altered Khrushchev's view of the American people. Although Khrushchev might have revised his alleged estimate of the American public had it enthusiastically supported a hard-line strategy, it is also plausible that had Kennedy committed American forces to the Bay of Pigs, he might have embroiled his administration in a politically divisive and militarily costly quagmire that only reinforced Khruschchev's view of Americans. Here again, no compelling political logic connects antecedent to consequent.

The complexity of the connecting principles underlying counterfactual arguments creates plenty of opportunities for running afoul of the cotenability standard (Goodman 1983; Elster 1978). Consider, for example, one part of Jon Elster's (1978) critique of Robert Fogel's (1964) classic counterfactual that "if the railroads had not existed, the American economy in the nineteenth century would have grown only slightly more slowly than it actually did." Elster argues that it is nonsensical to postulate as a supportive connecting principle that the internal combustion engine would have been invented earlier in the America without railroads because the postulate presupposes a theory of technical innovation that undercuts the original antecedent. If we have a theory of innovation that requires the invention of cars fifty years earlier, why does it not also require the invention of railroads? (Of course, Fogel's core counterfactual claim concerning the limited economic impact of railroads may still be correct even without speeding up the invention of automobiles. It rests on complex comparisons of the actual world with an elaborate counterfactual model of a nineteenth-century American economy that relied on waterways instead of railroads.)

In a similar spirit, Lars-Erik Cederman (Chapter 11) criticizes John Mueller's (1989) claim concerning the "irrelevance" of nuclear weapons. Mueller constructs a counterfactual non-nuclear scenario to demonstrate that nuclear weapons did not contribute to the postwar peace. Cederman notes that "the problem with Mueller's account is that he explicitly traces postwar history as it actually happened, including the Cuban missile crisis, while merely subtracting nuclear technology." This procedure illustrates the perils of superficially rewriting history. It is not at all clear that cotenability obtains between the counterfactual antecedent of a non-nuclear world and any connecting principle that posits the occurrence of the Cuban missile crisis in 1962. Why would the Soviets go to all the trouble of placing conventionally armed intermediate-range missiles in Cuba? Why take so large a risk for so small an advantage? There is something odd about the hypothetical world that Mueller created.

The two standards considered so far—logical clarity and cotenability—are helpful for screening out ambiguous and oxymoronic counterfactuals; purely formal (content-free) standards are not helpful, however, for screening out counterfactuals that invoke bizarre antecedents or connecting principles. What, for example, should we make of the suggestion that "if Napoleon had possessed a Stealth bomber, he would have won the Battle of Waterloo," or that "if Oswald did not shoot Kennedy, then someone else would have done so, because Kennedy was astrologically fated to die by assassination"? An adequate normative theory of counterfactual inference should give us principled grounds for rejecting conjectures of this sort. We see four ways of preempting such nonsense and we take up each in turn.

3. Consistency with Well-Established Historical Facts

Several scholars have proposed a "minimal-rewrite-of-history" rule designed to eliminate far-fetched counterfactuals that radically transform the temporal landscape (cf. Hawthorn 1991, 158; Weber 1949). They propose that, in principle, possible worlds should: (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. As noted earlier, these guidelines represent ground rules for assessing historical "possibility-hood." Operationally, investigators might agree to constrain conterfactual speculation in a host of more specific ways: by considering as antecedents only those policy options that participants themselves considered and (ideally) almost accepted, by giving extra weight to counterfactual antecedents that "undo" unusual events that appear to have made the decisive difference between the occurrence and nonoccurrence of the target event (and perhaps only the target event), by ruling out counterfactuals in which the antecedent and consequent are separated by such wide gaps of time that it is silly to suppose that all other things can be held equal, and by linking antecedent and consequent with connecting principles that are faithful to what we know about how people at the time thought and about the constraints within which people at the time had to work. This complex set of rules contains potential contradictions, but it does capture the flavor of most idiographic forms of counterfactual analysis (Chapter 3, Breslauer; Hart and Honoré 1959; Nash 1991).

Variants of the minimal-rewrite rule appear at several points in this volume. Scholars often invoke the rule to challenge or defend the legitimacy of considering certain counterfactual antecedents. For example, Lebow and Stein (Chapter 5) use this criterion to eliminate the "early warning" counterfactual that "had President Kennedy issued a timely warning in the spring of 1962, Khrushchev might not have sent missiles to Cuba." According to Lebow and Stein, the antecedent is implausible because it requires rewriting too much history. They note that "in April, before the conventional 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 antecedent, then, we have to rewrite history to alter both the political incentives and the evidence confronting the U.S. government. Using a similar standard, Khong (Chapter 4) argues for the plausibility of the antecedent in the counterfactual that "if Britain had confronted Hitler over Czechoslovakia, he would have backed down and World War II might have been avoided." According to Khong, the decisive factor in Britain's unwillingness to risk war at the time of Munich was neither the memory of World War I nor the unfavorable military balance, but Chamberlain's personal conviction that he could negotiate a diplomatic solution with Hitler. Khong finds historical evidence that influential politicians, including Eden, Cooper, and, of course, Churchill, favored a strong stance against Hitler as early as 1937. Hence, to the extent that any of these men could have been prime minister at the time of Munich, the antecedent becomes plausible.

Scholars also use the minimal-rewrite rule to assess the plausibility of connecting principles. For instance, Lebow and Stein (Chapter 5) assess Khrushchev's counterfactual claim that had the Soviet Union not deployed missiles in Cuba, the United States would have invaded the island. Lebow and Stein point to recently uncovered evidence that even before the missile deployment, no influential members of the Kennedy administration wanted to attack Cuba. The option had been considered but decisively rejected. Kennedy and Secretary of Defense McNamara had been impressed by Cuban popular support for Fidel Castro and the ability of the Cuban militia to overwhelm the invasion force at the Bay of Pigs. Revised intelligence estimates indicated that a successful invasion would require massive U.S. forces, which would have to remain in an occupational role for an indefinite period. Kennedy and McNamara were deterred by these costs and resolved not to attack unless there were dramatic political changes inside Cuba.

It is worth emphasizing that consistency with well-established historical facts may often be a necessary but is rarely a sufficient condition for establishing the plausibility of counterfactuals. As Breslauer (Chapter 3) notes, most counterfactual claims advanced in the Sovietological literature were consistent with historical evidence. He observes that "the problem was not invention of facts, but gaps in established bodies of facts, which allowed for multiple interpretation of the meanings of those pools of evidence."

4. Consistency with Well-Established Theoretical Laws

Just as we need historical and logical constraints on counterfactual reasoning, we also need theoretical constraints on the connecting principles we use to link antecedents and consequents. Otherwise, we cannot rule out counterfactuals that start from reasonable antecedents but end in far-fetched consequences by invoking preposterous principles of causality such as: "If Oswald had not shot Kennedy, then someone else would have done so, because Kennedy was astrologically fated to die by assassination," or "If North Korea had conquered South Korea in 1950, the economy of the South would have grown even more rapidly than it actually did because of the wisdom of the policy of self-sufficiency of the Great Leader Kim Il Sung."

Ideally, we could ground all counterfactual inferences in extensively validated scientific laws of the sort we drew upon in the match-lighting conditional. But do we have theoretical laws of comparable scope and power in the behavioral and social sciences? Some contributors, such as Lebow and Stein (Chapter 5), reject the notion that there are *any* well-established theories of international politics. They evaluate counterfactuals concerning the Cuban missile crisis by relying largely on case-specific political and historical standards. Other contributors, such as Kiser and Levi (Chapter 8), enthusiastically embrace deductive theory as a means of disciplining otherwise unruly "what-might-have-been" speculation.

The economic historian Robert Fogel is perhaps the preeminent advocate of the view that it is reasonable to rely on strong theory to fill in the missing counterfactual data points. Theory-guided counterfactuals are absolutely essential for assessing the economic impact of policies and technologies:

> The net effect of such things on development involves a comparison between what actually happened and what would have happened in the absence of the specified circumstance. However, since the counterfactual never occurred, it could not have been observed and hence is not recorded in historical documents. In order to determine what would have happened in the absence of a given circumstance, the economic historian needs a set of general statements (that is, a set of theories or

> model) that will enable him to deduce a counterfactual situation from institutions and relationships that actually existed. (Fogel 1964, 224)

In this view, counterfactual reasoning is a straightforward application of Hempel's (1965) covering law of historical explanation: "Counterfactual propositions [in quantitative economic history] are merely inferences from hypothetico-deductive models" (Fogel 1964).

Following the Hempel-Fogel neopositivist tradition, many game theorists, neoclassical economists, and structural realists display impressive confidence in their counterfactual claims. They know that if one changes the incentives confronting rational actors, those actors will quickly identify the new utility-maximizing course of action. If a currency is under- or overvalued, arbitrageurs will seize upon profit-making opportunities. If state regulations reward inefficiency and punish efficiency, aggregate economic output will fall. If a status quo power offers weak or incredible promises of extended deterrence to its allies, aggressors with much to gain and little to lose will strike. The calculus of rational action is not, however, the only theoretical logic that we can use to infer what would have happened in this or that counterfactual scenario. We can draw upon sociological theories that stress normative and institutional rules of fairness, cultural theories that stress group values and identifications, political theories of bureaucratic and interest-group competition, and cognitive theories of belief systems and bounded rationality.

Consider this sampling of the range of theories that political observers draw upon to "fill in" missing counterfactual data points:

(1) Keohane (1984) supports his claim that "if there were no international regimes, there would be less cooperation" by appealing to the Coase-Williamson tradition of institutional economics that stresses the role that institutions play in reducing the transaction costs of cooperation and in increasing the reputation costs of defection.

(2) Breslauer (Chapter 3) reviews the work of comparativists and area specialists who bolstered counterfactual claims about the causes of the Russian Revolution and the impact of Stalin's modernization policies by invoking theories of economic development.

(3) In their recent book, Lebow and Stein (1994) use Janis and Mann's (1977) psychological theory of decision making under stress to defend their claim that even if Kennedy had displayed more resolve at the Vienna summit, Khrushchev still would have deployed missiles in Cuba. Lebow and Stein argue that Khrushchev confronted a strategic dilemma that may well have induced a psychological state of defensive avoidance that, in turn, would have rendered Khrushchev insensitive to any plausible American signal of resolve.

(4) Kiser and Levi (Chapter 8) note how large classes of "agency" counterfactuals are ruled out by structural theories of revolution. If we view revolution as inevitable when certain structural preconditions are satisfied—intense international and demographic pressures on the state, fiscal crisis, deep divisions within the dominant class, and mass mobilization of discontented groups—then there is little point in contemplating counterfactuals that assign decisive roles to the actions of individuals. Within structuralist frameworks, it is impossible to undo the English, French, or Russian revolutions by simply positing wiser kings.

(5) Cederman (Chapter 11) draws on cartel theory to support his critique of neorealism and his claim that if defense-dominance had prevailed throughout history, there would have been less stability.

(6) Herrmann and Fischerkeller (Chapter 6) invoke Huth and Russett's theory of extended deterrence (a theory indigenous to political science, not an import) to argue that even if Truman had not threatened Stalin, the Soviets still would have withdrawn from Iran.

(7) Perhaps the most systematic use of theory to assess counterfactuals occurs in the chapters by Bueno de Mesquita and Weingast. These authors argue (among other things) that possible worlds become plausible only insofar as they are logically consistent with the equilibrium conditions of the game that captures the strategic interdependence obtaining between actual historical actors. For example, it does not make much sense to posit a hypothetical world in which both players cooperated in a single-round game if one or both of the players could have been much better off by defecting whenever the other player cooperated. Using this game-theoretic screening rule, it is possible to eliminate vast numbers of counterfactuals.

This overview is, however, disturbing. It suggests that each school of thought can foster its own favorite set of supporting counterfactuals (Chapter 12, Weber). Moreover, these schools of thought will sometimes prescribe contradictory rules for assessing counterfactuals. Where does this leave us? Consistency with *well-established* theory is a reasonable standard for gauging the plausibility of counterfactuals but we should expect disagreement about what counts as well-established theory in world politics. To prevent competing schools of thought from simply inventing counterfactuals of convenience, we need reality constraints. Counterfactuals must not only fit existing historical and statistical data (the emphasis in our third and fifth standards), they must stimulate testable predictions that hold up reasonably well against new data (the emphasis in our sixth standard, projectability).

5. Consistency with Well-Established Statistical Generalizations

In many contexts, we rely not on theoretical laws but on statistical generalizations to fill in "what would have happened if this rather than that event had occurred." One obvious form such reasoning takes is reliance on base rates and patterns of covariation. For instance, we might justify the counterfactual, "If Bill Clinton had lost the presidential election of 1992, he would have been disappointed," by observing in a two-by-two contingency table that when people fail to achieve a goal for which they have worked long and hard, the overwhelming majority experience disappointment, but when people do achieve their goals, there is markedly less disappointment.

The "discovery" is hardly startling; more startling, however, is the strong stand that some scholars take on both the *necessity* and *sufficiency* of statistical justification for assessing counterfactual claims. In his commentary, Dawes, for instance, treats statistical evidence as trump when he declares that counterfactual inferences are justified *if and only if* they are embedded in a system of statistical contingency for which we have reasonable evidence. He offers an intriguing example:

> Suppose that someone is required to wager her entire wealth on a single roll of a pair of fair dice. Her wealth will be doubled if she wins the bet; if she loses, she will be bankrupt. Her choice is to bet either for or against a roll of snake eyes. Being wise, she bets against snake eyes. The dice are rolled and they come up snake eyes. She loses. She is bankrupt. Is it normatively valid to state that "if only" she had bet on snake eyes, she would have won? Well, it is true that she bet against snake eyes and that she lost. But does the "if only" add anything to the analysis? I suggest the answer is no. But suppose she had bet on snake eyes and lost. Here, I suggest that the regretful counterfactual inference that "if only" she had bet against snake eyes she would have won is normatively justified. Why? Because the odds are thirty-five to one against snake eyes, and those odds justify the expectation that she would have won had she bet against snake eyes. It's an expectation; one can insert "probably won" if one wishes.

The plausibility of the snake-eyes counterfactual hinges on the aim of inquiry. From an historical point of view, the counterfactual is plausible. We do not have to rewrite much history to reach a hypothetical world in which the woman threw the dice slightly differently and won her bet. The bet against snake eyes is an easily imagined and easily reversed cause of bankruptcy. To statisticians of both the Bayesian and frequentist schools, however, the counterfactual is implausible because it posits so unlikely an outcome. We certainly do not want people drawing the lesson from history that it is a good idea to risk their fortunes on long-shot wagers.[3]

One need not, of course, accept the radical epistemological argument of Dawes to agree with the more moderate mainstream view that canons of sound statistical reasoning should constrain our judgments of counterfactuals (King, Keohane, and Verba 1994) and, indeed, that we should be alert to the psychological fact that people are flawed intuitive statisticians who fall prey to various biases in detecting and using covariation data. The experimental literature warns us that people often draw inappropriately strong conclusions from observing only the cause-present/effect-present cell of contingency tables and see strong relationships between variables that they expect to be correlated but are in reality only weakly correlated (Nisbett and Ross 1980). A good start in implementing our fifth criterion would be simply to improve the accuracy of intuitive estimates of covariation, with special attention to sensitizing people to the problem of missing counterfactual data. Accurate covariation estimates would, however, be just the beginning and would not protect us from accepting many false counterfactual claims (Type I errors) and rejecting many true ones (Type II errors). We also need to beware of biases produced by nonrandom selection, confounding variables, and omitted variables, as well as a host of other familiar obstacles to meaningful statistical inference (King, Keohane, and Verba 1994). These statistical issues play an especially prominent role in Russett's chapter, which grapples with the controversial counterfactual that if all states in the twentieth century had been democratic, there would have been markedly fewer wars. Skeptics of the democratic-peace hypothesis challenge this counterfactual on various grounds, including the inadequacy of the available statistical samples (too few democracies, too few wars, and too truncated a range of time), the inadequacy of the operational definitions of democracy and war (the self-serving suppleness of certain judgment calls), and—perhaps most important—the collinearity problems created by confounding variables that, once controlled for in regressions, may "explain away" the democracy effect. Russett responds by rebutting these objections, in the process illustrating the enormous overlap between traditional procedures for hypothesis testing and

[3] Dawes's snake-eyes problem bears a deep resemblance to Newcomb's paradox, which pits statistical intuition against causal intuition through an ingenious thought experiment that calls upon us to imagine a being who has demonstrated a phenomenal capacity to make accurate predictions ($R^2 = 1.0$) and who has asked us to make a choice involving two boxes, *B*1 and *B*2. *B*1 contains $1,000; *B*2 contains either $1,000,000 or nothing. Our choice is between two actions: (1) taking what is in both boxes; (2) taking only what is in the second box. Furthermore, we know, and the being knows we know, and so on, that if the being predicts that we will take what is in both boxes, he will not put the $1 million in the second box; if the being predicts we will take only what is in the second box, he will put the $1 million in the second box. The rules are straightforward. First the being makes his prediction; then he puts the $1 million in the second box or not, according to his prediction; then we make our choice.

The problem is paradoxical because powerful epistemic intuitions push us in opposite directions (Nozick 1993). Statistical intuition tells us that if we take what is in both boxes, the being almost certainly will have predicted this choice and will not have put the $1 million in the second box, whereas if we take only what is in the second box we will almost certainly get $1 million. Therefore, we should take only what is in the second box. Causal intuition tells us, however, that the being has made his prediction and has already either put the $1 million into the second box or has not. That fact cannot be undone. Therefore, we will receive more money, $1,000 more, by taking what is in both boxes.

Thought experiments of this form provide a useful means of clarifying our intuitions about how to resolve clashes between epistemic standards.

those for evaluating an important category of counterfactual (our ideal type 2, the nomothetic).

Statistical tests of counterfactual plausibility also play a pivotal role in the chapters by game theorists. Weingast's notion of comparative statistics reminds us of the need to build appropriate time lags into our assessments of covariation. And Bueno de Mesquita's work on medieval church-state relations reminds us of the need for probabilistic tests of hypotheses concerning mixed-strategy equilibria.

6. Projectability

Theory evaluation and counterfactual evaluation are inextricably entangled. Sound counterfactuals require sound theories that provide the lawlike generalizations that fill in the missing data points in our thought experiments. How can we judge, however, whether these lawlike generalizations are robust enough to support counterfactual inferences? Here Nelson Goodman's (1983) concept of projectability is helpful. Goodman draws a sharp distinction between coincidental generalizations that just happen to be true at a particular time and place (and are therefore unprojectable) and robustly lawlike generalizations that hold up over a range of circumstances and permit projection into the past and future. An example of a merely coincidental generalization is "All the coins in my pocket yesterday were silver." Nothing follows from this observation—certainly not "If this penny were in my pocket yesterday, it would be silver." The counterfactual fails because "if this penny were in my pocket yesterday," we would simply assume that the original generalization—"all the coins in my pocket yesterday were silver"—was false. By contrast, a robustly lawlike generalization—such as that oxygen is a necessary but not sufficient condition for fire—inspires confidence when we move either backward in time (if there had been no oxygen, the Great Fire of London would not have occurred) or forward in time (if we cut off any future fire's source of oxygen, the fire will expire).

Most social-science generalizations, of course, qualify as neither merely coincidental nor robustly lawlike; they take the form of either contingent generalizations (under this set of boundary conditions, *x* causes *y*; under that set, *x* causes *z*) or statistical generalizations (*x* increases or decreases the likelihood of *y*) or contingent statistical generalizations (cf. George and Smoke 1974; George 1993). From Goodman's perspective, however, whether the generalization is bounded or unbounded by moderator variables and whether the generalization is deterministic or probabilistic, it is subject to the same acid test of scientific legitimacy: namely, its projectability or its ability to predict what will happen in new, hitherto unobserved cases. The same causal principles that allow us to retrodict the past should allow us to predict the future. Indeed, the strong Popperian form of this argument asserts that we should take counterfactual claims seriously if and only if the lawlike generalizations supporting the claims yield falsifiable forecasts. We see this classic philosophical argument resurfacing in the recent methodological advice of King, Keohane, and Verba (1994), who urge scholars to search aggressively for the observable implications of their causal constructs by regularly asking themselves, "If my argument is correct, what else should be true?" Counterfactuals that are devoid of testable implications in the actual world leave us marooned in hypothetical worlds of our own subjective making. Projectability, from this vantage point, stands as the preeminent criterion for judging the value of counterfactual speculation.

Perhaps not coincidentally, the most outspoken advocates of the projectability standard in this volume tend to be the most nomothetic in their overall approach to social science. Bueno de Mesquita and Weingast are not content with post-hoc exercises in which they fit game-theoretic models to data; they derive testable implications from their models and show how those predictions can be statistically or historically disconfirmed. In a similar vein, Russett is not satisfied with showing that the democratic-peace hypothesis captures an intriguing regularity in the brief slice of history that we call modern; he seeks out alternative data sources—Greek city-states and tribal societies—into which we can project the hypothesis. In his computer simulations, Cederman explores the replicability and robustness of his counterintuitive result that defense-dominance increases rather than decreases the likelihood of the emergence of hegemons. And in their critique of structuralist theories of revolution, Kiser and Levi raise the suspicion that structuralist theories—with their emphasis on complex conjunctions of preconditions—are ultimately exercises in post-hoc data-fitting that will never pass the projectability test.[4]

[4] The strong version of the "projectability" argument treats backward and forward reasoning in time as symmetrical. There are good reasons, however, for suspecting that even when we can construct compelling explanations of the past, we will often do a terrible job of explaining the future (Dawes 1993; Chapter 2, Fearon). When we look back into the past from the present, we occupy a privileged but also easily abused position. We know which one of the many futures that were once possible has actually occurred. With the benefit of this retrospective knowledge, it becomes relatively easy to find antecedents that depict the consequence as the inevitable result of some "inexorable" causal process. Yet we risk capitalizing, indeed massively capitalizing, on chance. By contrast, when we look forward into the future, we cannot avoid the complexity and indeterminacy of possible relationships among antecedents and consequences. We can draw upon our knowledge of past causal relationships to anticipate the future, but are often disappointed by the results. The causes we identified retrospectively for a class of consequences prove to be anemic predictors of the same class in the future. Dawes (1993) illustrates this argument with the crash of a passenger airplane into a parked truck on a runway under repair at Mexico City airport on October 31, 1979. The FAA flight investigators easily constructed a causal story for the outcome that invoked such plausible antecedents as poor weather, smog, pilot fatigue, radio malfunction, cryptic communication by traffic controllers, and stress. But

Psychological Perspectives on Counterfactual Reasoning

Up to this juncture, we have focused on normative perspectives on counterfactual reasoning—on the criteria that people should use to generate and judge counterfactual arguments. We now turn to psychological perspectives. There is a thriving research literature in both cognitive psychology (Commentary 2, Olson, Roese, and Deibert) and linguistics (Commentary 1, Turner) on how people actually generate and judge counterfactual claims. These normative and psychological arguments should not, of course, be viewed as two self-contained, hermetically sealed domains of discourse. The psychological literature highlights a host of determinants of spontaneous counterfactual reasoning that raise serious questions about the reliability and validity of counterfactual thought experiments in world politics. Indeed, when the topic is thought experiments, it is hard to say at what point epistemology and methodology end and psychology begins.

From a broadly psychological perspective, it is difficult to imagine avoiding serious bias in thought experiments. Bias can creep into every stage of this inherently subjective process, from the initial selection of antecedents (for "mental manipulation") to the evaluation of connecting principles to the willingness to entertain counterarguments and alternative scenarios. Bias appears inevitable, in part because of the cognitive limitations and motivational inclinations of the thinker in whose mind the thought experiment "runs," and in part because of the extraordinary complexity and ambiguity of the task. The population of past events from which one can draw counterfactual antecedents is effectively infinite, from the flapping of butterfly wings to the "structural polarity" of the international system. And the task of assessing what would have happened in these hypothetical worlds (to which no one has access) is obviously highly subjective. Consider the potential for epistemic mischief.

Cognitive Biases

A useful starting point is the principle of bounded rationality (Simon 1957). People, it is now widely conceded, are limited-capacity information processors who rely on low-effort strategies to simplify an otherwise intolerably complex world (Kahneman, Slovic, and Tversky 1982). The price of cognitive economy is, however, steep: increased susceptibility to systematic biases and errors. We itemize several ways in which reliance on simplifying strategies might distort the conclusions we draw from counterfactual thought experiments.

(1) *What gets mutated?* This is the ground floor for the entry of psychological bias into thought experiments. As Kahneman and Miller (1986) argue in their influential norm theory, the human perceptual apparatus is attuned to notice change. The more abrupt or discontinuous the change, the more likely people are to notice it, to try to explain it, and to generate counterfactual scenarios in which they "mutate" the departure from normality to the more customary and expected default value (Commentary 2, Olson, Roese, and Deibert). For example, experimental subjects generate more "if only" thoughts and experience more regret upon learning that the victim of a traffic accident had departed from her regular route to the office than they do upon learning that the accident victim had adhered to her regular route. It is easier to "mentally undo" accidents or indeed other events that constitute deviations from the routine.

Translating this well-replicated finding from the experimental literature into the realm of world politics is no simple exercise. This volume, however, contains much evidence that departures from normality or the status quo do indeed attract especially vigorous counterfactual speculation. These departures can take diverse forms, including leadership transitions (Chapter 3, Breslauer; Chapter 6, Herrmann and Fischerkeller), revolutions (Chapter 8, Kiser and Levi), assassinations (Chapter 3, Breslauer), and unusually intense policy debates in which the argument might easily have gone either way (Chapter 4, Khong; Chapter 5, Lebow and Stein). Routine events fade into the perceptual background and are rarely selected for mental manipulation in thought experiments. Of course, not all social scientists conform to the predictions of norm theory. Nomothetic investigators often invoke background conditions that change almost imperceptibly slowly (such as the size of the middle class in early twentieth-century Russia or the "polarity" of the pre–World War I balance of power). And chaos and complexity theorists specialize in demonstrating the sensitive dependence of major outcomes on minor background conditions, such as the flapping of butterfly wings. Norm theory fits idiographic better than nomothetic "counterfactualizing."

(2) *Once constructed, which counterfactual scenarios are judged plausible?* The simplifying strategies that people use to impose cognitive order carry a price tag. These strategies can tilt the playing field (arguably unfairly) in favor of certain counterfactuals over others. Consider the much discussed trilogy of judgmental heuristics: anchoring, availability, and representativeness (Tversky and Kahneman 1974). The anchoring heuristic could lead people to be too quick to dismiss scenarios about hypothetical worlds

the causal variables identified in that story will probably not help FAA investigators to predict future crashes. For example, pilots are often tired, but rarely crash, and many crashes occur when pilots are well rested. This argument suggests that counterfactuals in world politics often fail the projectability test not because the underlying claims are false, but rather because there are such complex interactions among causal variables and so much potential for randomly distributed small causes to be amplified into large effects (a point reminiscent of chaos theory). In this view, projectability is most likely to break down for explanations of low-likelihood and low-frequency events (exactly the sorts of surprising events that, psychologists argue, are most likely to attract counterfactual speculation).

that deviate dramatically from the perceptual anchor of the actual world with which they are already so familiar (making it difficult to appreciate the arbitrariness of the status quo); the availability heuristic could lead people to be too quick to embrace vivid, easily imaginable scenarios that link all the component events into a compelling story (even though the compound probability of all the narrative's components taken together is vanishingly small); the representativeness heuristic could lead people to be too slow to concede plausibility to counterfactuals that posit dramatic nonlinearities in cause-effect relations (making it difficult to appreciate that small causes can sometimes produce big effects and vice versa).

Perhaps the most lethal threat to the validity of counterfactual thought experiments comes, however, from theory-driven thinking. We have already noted that counterfactual reasoning will inevitably be theory-driven to some degree. Indeed, we treated "consistency with well-established theory" as a defining feature of sound counterfactual reasoning. But the cognitive perspective leads us to be suspicious of people's capacity to apply standards of evidence and proof in an evenhanded fashion (Nisbett and Ross 1980; Fiske and Taylor 1991). People often succumb to the temptation of applying strong tests to dissonant arguments and weak tests to consonant ones—a temptation that may be especially pronounced when the arguments invoke possible worlds that no one can ever enter and that can never be decisively disconfirmed. The perceived plausibility of a counterfactual hinges on how hard one looks for shortcomings. Few counterfactual arguments will not have points of vulnerability when we subject their antecedents and connecting principles to close scrutiny. As a result, we are much more likely to recognize the collapse of cotenability in our opponents arguments than in our own—a recurring theme in several chapters.

The cognitive perspective also leads us to be suspicious of people's capacity to transcend (avoid contamination by) outcome knowledge. As theory-driven thinkers, people automatically try to assimilate "what happened" to some prior knowledge structure or schema that specifies cause-effect relationships for events of that type (Fischhoff 1975; Hawkins and Hastie 1990). The result is a deep, and arguably unjustifiable, asymmetry between backward and forward reasoning in time. On average, political experts see fewer possible pasts than they do possible futures (Tetlock 1994). When they look backward in time, they mobilize their finite mental resources to explain the one outcome of the many possible outcomes that actually occurred, selectively recruiting the most plausible (theory-consistent) antecedents that will allow them to tell a good causal story for that outcome (Commentary 3, Dawes; Chapter 12, Weber). By contrast, when experts look forward into the future, they are typically unsure of what will happen. In part for contingency planning and in part to avoid the embarrassment of making blatantly wrong forecasts, experts often survey in a reasonably open-minded way the panoply of possibilities and conditions for their occurrence. This cognitive analysis helps us to explain an otherwise paradoxical pattern in expert judgment: bold counterfactuals and timid forecasts. Experts often assert that they know what would have happened in the past but modestly demur on what will happen in the future.

Motivational Biases

Thus far, we have focused on only cognitive sources of bias. People are not, of course, just information-processing devices; they are animated by wishes, hopes, and fears that shape their perceptions of what might or could or should have been (Commentary 2, Olson, Roese, and Deibert). These emotional needs can take many, sometimes conflicting forms (Tetlock and Levi 1982). Consider the following possibilities suggested by the psychological literature:

(1) *Needs for predictability and controllability*: On the one hand, people might allow their desire to believe that the world is fundamentally predictable to rule out butterfly-effect counterfactuals, which imply that, no matter how hard we try, it is in principle impossible to anticipate the future because so much hinges on small causes that are beyond our measurement grasp. On the other hand, people might allow their desire to believe that the world is controllable to rule out "inevitability" counterfactuals, which imply that, no matter what people do, our fates are ultimately under the sway of powerful geopolitical, macroeconomic, and technological forces beyond individual mastery. Indeed, this psychological need to believe in a "controllable world" may lie at the heart of Kissinger's conversion from his belief as an academic observer in deterministic arguments that minimized policy makers' latitude to influence events, to his world view as a policy maker which assigned a much more prominent role to individual human beings who could be persuaded—through the right combination of arguments and inducements—to change their minds (Chapter 3, Breslauer; Kissinger 1993).

(2) *Needs to avoid blame and to claim credit*: On the one hand, people might allow their desire to avoid blame for bad outcomes to override their desires for predictability and control. In such cases, people will argue that they should not be blamed for having failed to foresee the unforeseeable or having failed to control the uncontrollable. On the other hand, people might allow the desire to claim credit for good outcomes to enhance the plausibility of counterfactuals that take the form "had it not been for my superior predictive ability and courageous willingness to act on the basis of that insight, this good outcome would never have occurred."

(3) *Needs for consolation and inspiration*: People might use "downward" counterfactuals to comfort and console themselves ("Things may not be

great, but think how bad things could have been if *x* or *y* had occurred") or "upward" counterfactuals to inspire greater effort ("Do not be complacent about the present, think how good things could have been and, by implication, could yet become").

(4) *Need for cognitive consistency*: The well-documented aversion to imbalanced or dissonant couplings of events should motivate people to rule out counterfactuals that link bad causes (like Stalin) to good outcomes (like accelerated economic growth) or that link good causes (like foreign aid) to bad outcomes (like increased dependency and corruption of recipient regimes). Pressures for cognitive consistency should also motivate people to defend "core beliefs." For example, people who believe that "evil is avoidable" should be strongly motivated to generate counterfactuals that undo moral catastrophes. But people of different political persuasions may define moral catastrophe differently. For many conservatives, the root of evil in Soviet history goes straight back to the Bolshevik revolution of 1917 (which should be a focal point of "if only" speculation); for many social democrats, a noble socialist experiment was corrupted by Stalinist tyranny (which should be a focal point of "if only" speculation). These predictions fare reasonably well against the evidence (Chaper 3, Breslauer).

The list is a lengthy and unparsimonious one. Here we simply want to add that there are always two levels at which motives may influence counterfactual reasoning: private thought (affecting what we truly believe) and public posturing (affecting what we say we believe and want to induce others to believe). Most psychologists think that the motives listed here do indeed shape privately held plausibility judgments of counterfactuals, but few would deny that public impression management is also at work (Tetlock and Manstead 1985)—a judgment with which most of our contributors seem to concur. This volume contains suggestive evidence that the closer we get to prescriptive policy debates, the greater the temptation to use counterfactual arguments as rhetorical tools to justify either what one plans to do or has already done (Chapter 3, Breslauer; Chapter 5, Lebow and Stein; Chapter 6, Herrmann and Fischerkeller).

How should we respond to this extended list of cognitive and motivational threats to the validity of thought experiments? We urge a middle-ground response between arrogance and despair. The arrogant response is, of course, to argue that although mere mortals may fall prey to these biases, serious professionals are surely immune. We judge this response arrogant because there is already abundant evidence that a wide range of professionals working on important tasks are susceptible to many of the effects discussed here (Dawes 1991). Moreover, some of our contributors report evidence of certain hypothesized biases in the scholarly literature (Chapter 3, Breslauer; Chapter 8, Kiser and Levi). The despairing response is, of course, to argue that the biases identified by psychologists are inevitable—that they have hopelessly contaminated all counterfactual arguments advanced thus far and will contaminate all counterfactuals advanced into the foreseeable future. The former response is too dismissive of the psychological literature; the latter takes it too literally. The appropriate response in our view is to acknowledge cognitive biases and to make good-faith efforts to hold each other accountable to standards of evidence (such as those sketched in this chapter) that check the most serious and pervasive of these biases (Tetlock 1992b). All research methods are subject to contamination and misinterpretation; it is only prudent to beware of potential biases in the most subjective of all methods of inquiry, the counterfactual thought experiment.

Conclusion

There is something about the topic of counterfactual thought experiments in world politics that makes people feel a bit uneasy, even defensive. To be blunt, it feels like epistemological slumming. As social scientists, we are all too familiar with the prestige hierarchy for methods of drawing causal inference. At the top of the scientific pecking order is experimentation in which we can manipulate hypothesized causes and then either hold everything else constant or randomize extraneous influences across treatment conditions. Experimental control of this sort is obviously out of the question for most questions in world politics. We cannot rerun the tape of history: splicing a Gorbachev in or out, delaying or accelerating key technological development, or tinkering with this or that aspect of macroeconomic policy.

Social scientists often resort to statistical control when experimentation is ethically or practically problematic. But statistical arguments themselves often rest on counterfactual assumptions (Fearon 1991) and are, in any case, extraordinarily difficult to make for many issues that loom large in security debates. For example, what kind of regression or time series analyses will allow us to estimate the causal contribution of nuclear weapons to the "long peace" between the United States and Soviet Union between 1945 and 1991? There are simply too many confounding variables—a problem we can alleviate but not eliminate through judicious selection of comparison cases and meticulous process-tracing of decision-making protocols.

So where does that leave us? Probably still feeling uneasy: we seem to be stuck with quite literally a third-rate method, counterfactual thought experimentation. The control groups exist—if indeed "exist" is the right word—in the imaginations of political analysts who are left with the daunting task of reconstructing how history would have unfolded if causal variables of the past had taken on different values from the ones they actually did. The whole exercise starts to look hopelessly subjective, circular, and nonfalsifiable. What is to stop us from simply inventing counterfactual outcomes that 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).