Expert Political Judgment

HOW GOOD IS IT? HOW CAN WE KNOW?

Philip E. Tetlock

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Contents

Acknowledgments ix

Preface xi

CHAPTER 1
Quantifying the Unquantifiable 1

CHAPTER 2
The Ego-deflating Challenge of Radical Skepticism 25

CHAPTER 3
Knowing the Limits of One's Knowledge: Foxes Have Better Calibration and Discrimination Scores than Hedgehogs 67

CHAPTER 4
Honoring Reputational Bets: Foxes Are Better Bayesians than Hedgehogs 121

CHAPTER 5
Contemplating Counterfactuals: Foxes Are More Willing than Hedgehogs to Entertain Self-subversive Scenarios 144

CHAPTER 6
The Hedgehogs Strike Back 164

CHAPTER 7
Are We Open-minded Enough to Acknowledge the Limits of Open-mindedness? 189

CHAPTER 8
Exploring the Limits on Objectivity and Accountability 216

Methodological Appendix 239

Technical Appendix
Phillip Rescober and Philip E. Tetlock 273

Index 313

CHAPTER 1

Quantifying the Unquantifiable

> I do not pretend to start with precise questions. I do not think you can start with anything precise. You have to achieve such precision as you can, as you go along.
>
> —BERTRAND RUSSELL

EVERY DAY, countless experts offer innumerable opinions in a dizzying array of forums. Cynics groan that expert communities seem ready at hand for virtually any issue in the political spotlight—communities from which governments or their critics can mobilize platoons of pundits to make prepackaged cases on a moment's notice.

Although there is nothing odd about experts playing prominent roles in debates, it is odd to keep score, to track expert performance against explicit benchmarks of accuracy and rigor. And that is what I have struggled to do in twenty years of research of soliciting and scoring experts' judgments on a wide range of issues. The key term is "struggled." For, if it were easy to set standards for judging judgment that would be honored across the opinion spectrum and not glibly dismissed as another sneaky effort to seize the high ground for a favorite cause, someone would have patented the process long ago.

The current squabble over "intelligence failures" preceding the American invasion of Iraq is the latest illustration of why some esteemed colleagues doubted the feasibility of this project all along and why I felt it essential to push forward anyway. As I write, supporters of the invasion are on the defensive: their boldest predictions of weapons of mass destruction and of minimal resistance have not been borne out.

But are hawks under an obligation—the debating equivalent of Marquis of Queensbury rules—to concede they were wrong? The majority are defiant. Some say they will yet be proved right: weapons will be found—so, be patient—or that Baathists snuck the weapons into Syria—so, broaden the search. Others concede that yes, we overestimated Saddam's arsenal, but we made the right mistake. Given what we knew back then—the fragmentary but ominous indicators of Saddam's intentions—it was prudent to over- rather than underestimate him. Yet others argue that ends justify means: removing Saddam will yield enormous long-term

benefits if we just stay the course. The know-it-all doves display a double failure of moral imagination. Looking back, they do not see how terribly things would have turned out in the counterfactual world in which Saddam remained ensconced in power (and France wielded de facto veto power over American security policy). Looking forward, they do not see how wonderfully things will turn out: freedom, peace, and prosperity flourishing in lieu of tyranny, war, and misery.[1]

The belief system defenses deployed in the Iraq debate bear suspicious similarities to those deployed in other controversies sprinkled throughout this book. But documenting defenses, and the fierce conviction behind them, serves a deeper purpose. It highlights why, if we want to stop running into ideological impasses rooted in each side's insistence on scoring its own performance, we need to start thinking more deeply about how we think. We need methods of calibrating expert performance that transcend partisan bickering and check our species' deep-rooted penchant for self-justification.[2]

The next two sections of this chapter wrestle with the complexities of the process of setting standards for judging judgment. The final section previews what we discover when we apply these standards to experts in the field, asking them to predict outcomes around the world and to comment on their own and rivals' successes and failures. These regional forecasting exercises generate winners and losers, but they are not clustered along the lines that partisans of the left or right, or of fashionable academic schools of thought, expected. *What* experts think matters far less than *how* they think. If we want realistic odds on what will happen next, coupled to a willingness to admit mistakes, we are better off turning to experts who embody the intellectual traits of Isaiah Berlin's prototypical fox—those who "know many little things," draw from an eclectic array of traditions, and accept ambiguity and contradiction as inevitable features of life—than we are turning to Berlin's hedgehogs—those who "know one big thing," toil devotedly within one tradition, and reach for formulaic solutions to ill-defined problems.[3] The net result is a double irony: a perversely inverse relationship between my prime exhibit indicators of good judgment and the qualities the media prizes in pundits—the tenacity required to prevail in ideological combat—and the qualities

[1] For a passionate affirmation of these defenses, see W. Safire, "The New Groupthink," *New York Times*, July 14, 2004, A27.

[2] The characterization of human beings as rationalizing rather than rational animals is as old as Aristotle and as new as experimental social psychology. See Z. Kunda, *Social Cognition: Making Sense of People* (Boston: MIT Press, 1999).

[3] I. Berlin, "The Hedgehog and the Fox," in *The Proper Study of Mankind* (New York: Farrar, Straus & Giroux, 1997), 436–98. Berlin traces the distinction—via Erasmus—2,600 years to a shadowy source on the edge of recorded Greek history: the soldier-poet Archilocus. The metaphorical meaning oscillates over time, but it never strays far from eclectic cunning (foxes) and dogged persistence (hedgehogs).

science prizes in scientists—the tenacity required to reduce superficial complexity to underlying simplicity.

Here Lurk (the Social Science Equivalent of) Dragons

It is a curious thing. Almost all of us think we possess *it* in healthy measure. Many of us think we are so blessed that we have an obligation to share *it*. But even the savvy professionals recruited from academia, government, and think tanks to participate in the studies collected here have a struggle defining *it*. When pressed for a precise answer, a disconcerting number fell back on Potter Stewart's famous definition of pornography: "I know it when I see *it*." And, of those participants who ventured beyond the transparently tautological, a goodly number offered definitions that were in deep, even irreconcilable, conflict. However we set up the spectrum of opinion—liberals versus conservatives, realists versus idealists, doomsters versus boomsters—we found little agreement on either who had *it* or what *it* was.

The elusive *it* is good political judgment. And some reviewers warned that, of all the domains I could have chosen—many, like medicine or finance, endowed with incontrovertible criteria for assessing accuracy—I showed suspect scientific judgment in choosing good political judgment. In their view, I could scarcely have chosen a topic more hopelessly subjective and less suitable for scientific analysis. Future professional gatekeepers should do a better job stopping scientific interlopers, such as the author, from wasting everyone's time—perhaps by posting the admonitory sign that medieval mapmakers used to stop explorers from sailing off the earth: *hic sunt dragones*.

This "relativist" challenge strikes at the conceptual heart of this project. For, if the challenge in its strongest form is right, all that follows is for naught. Strong relativism stipulates an obligation to judge each worldview within the framework of its own assumptions about the world—an obligation that theorists ground in arguments that stress the inappropriateness of imposing one group's standards of rationality on other groups.[4] Regardless of precise rationale, this doctrine imposes a blanket ban on all

[4] Extreme relativism may be a mix of anthropological and epistemological posturing. But prominent scholars have advanced strong "incommensurability arguments" that claim clashing worldviews entail such different standards of evidence as to make mutual comprehension impossible. In philosophy of science: P. Feyerabend, *Against Method: Outline of an Anarchistic Theory of Knowledge* (London: Humanities Press, 1975). In moral theory, A. MacIntyre, *Whose Justice? Which Rationality?* (London: Duckworth, 1988). Such arguments carry strong implications for how to do research. We should adopt a nonjudgmental approach to judgment, one limited to compiling colorful ethnographic catalogs of the odd ideas that have prevailed at different times and places.

efforts to hold advocates of different worldviews accountable to common norms for judging judgment. We are barred from even the most obvious observations: from pointing out that forecasters are better advised to use econometric models than astrological charts or from noting the paucity of evidence for Herr Hitler's "theory" of Aryan supremacy or Comrade Kim Il Sung's juche "theory" of economic development.

Exasperation is an understandable response to extreme relativism. Indeed, it was exasperation that, two and a half centuries ago, drove Samuel Johnson to dismiss the metaphysical doctrines of Bishop Berkeley by kicking a stone and declaring, "I refute him thus." In this spirit, we might crankily ask what makes political judgment so special. Why should political observers be insulated from the standards of accuracy and rigor that we demand of professionals in other lines of work?

But we err if we shut out more nuanced forms of relativism. For, in key respects, political judgment is especially problematic. The root of the problem is not just the variety of viewpoints. It is the difficulty that advocates have pinning each other down in debate. When partisans disagree over free trade or arms control or foreign aid, the disagreements hinge on more than easily ascertained claims about trade deficits or missile counts or leaky transfer buckets. The disputes also hinge on hard-to-refute counterfactual claims about what would have happened if we had taken different policy paths and on impossible-to-refute moral claims about the types of people we should aspire to be—all claims that partisans can use to fortify their positions against falsification. Without retreating into full-blown relativism, we need to recognize that political belief systems are at continual risk of evolving into self-perpetuating worldviews, with their own self-serving criteria for judging judgment and keeping score, their own stocks of favorite historical analogies, and their own pantheons of heroes and villains.

We get a clear picture of how murky things can get when we explore the difficulties that even thoughtful observers run into when they try (as they have since Thucydides) to appraise the quality of judgment displayed by leaders at critical junctures in history. This vast case study literature underscores—in scores of ways—how wrong Johnsonian stone-kickers are if they insist that demonstrating defective judgment is a straightforward "I refute him thus" exercise.[5] To make compelling indictments of political judgment—ones that will move more than one's ideological soul

[5] For excellent compilations, and analyses, of such arguments, see R. Jervis, *Perception and Misperception in International Politics* (Princeton, NJ: Princeton University Press, 1976); R. E. Neustadt and E. R. May, *Thinking in Time* (New York: Free Press, 1986); Y. Vertzberger, *The World in Their Minds* (Stanford, CA: Stanford University Press, 1990); Y. F. Khong, *Analogies at War* (Princeton, NJ: Princeton University Press, 1993); B. W. Jentleson, ed., *Opportunities Missed, Opportunities Seized: Preventive Diplomacy in the*

mates—case study investigators must show not only that decision makers sized up the situation incorrectly but also that, as a result, they put us on a manifestly suboptimal path relative to what was once possible, and they could have avoided these mistakes if they had performed due diligence in analyzing the available information.

These value-laden "counterfactual" and "decision-process" judgment calls create opportunities for subjectivity to seep into historical assessments of even exhaustively scrutinized cases. Consider four examples of the potential for partisan mischief:

a. How confident can we now be—sixty years later and after all records have been declassified—that Harry Truman was right to drop atomic bombs on Japan in August 1945? This question still polarizes observers, in part, because their answers hinge on guesses about how quickly Japan would have surrendered if its officials had been invited to witness a demonstration blast; in part, because their answers hinge on values—the moral weight we place on American versus Japanese lives and on whether we deem death by nuclear incineration or radiation to be worse than death by other means; and, in part, because their answers hinge on murky "process" judgments—whether Truman shrewdly surmised that he had passed the point of diminishing returns for further deliberation or whether he acted impulsively and should have heard out more points of view.[6]
b. How confident can we now be—forty years later—that the Kennedy administration handled the Cuban missile crisis with consummate skill, striking the perfect blend of firmness to force the withdrawal of Soviet missiles and of reassurance to forestall escalation into war? Our answers hinge not only on our risk tolerance but also on our hunches about whether Kennedy was just lucky to have avoided dramatic escalation (critics on the left argue that he played a perilous game of brinkmanship) or about whether Kennedy bollixed an opportunity to eliminate the Castro regime and destabilize the Soviet empire (critics on the right argue that he gave up more than he should have).[7]

Post–Cold War World (Lanham, MD: Rowman & Littlefield, 1999); F. I. Greenstein, *The Presidential Difference: Leadership Styles from FDR to Clinton* (New York: Free Press, 2000); D. W. Larson and S. A. Renshon, *Good Judgment in Foreign Policy* (Lanham, MD: Rowman & Littlefield, 2003).

[6] D. McCullough, *Truman* (New York: Simon & Schuster, 1992); B. J. Bernstein, "The Atomic Bombing Reconsidered," *Foreign Affairs* 74 (1995): 147.

[7] D. Welch and J. Blight, "The Eleventh Hour of the Cuban Missile Crisis: An Introduction to the ExComm Tapes," *International Security* 12 (1987/88): 5–92; S. Stern, "Source Material: The 1997 Published Transcripts of the JFK Cuban Missile Crisis Tapes: Too Good to be True?" *Presidential Studies Quarterly* 3 (1997): 586–93.

c. How confident can we now be—twenty years later—that Reagan's admirers have gotten it right and the Star Wars initiative was a stroke of genius, an end run around the bureaucracy that destabilized the Soviet empire and hastened the resolution of the cold war? Or that Reagan's detractors have gotten it right and the initiative was the foolish whim of a man already descending into senility, a whim that wasted billions of dollars and that could have triggered a ferocious escalation of the cold war? Our answers hinge on inevitably speculative judgments of how history would have unfolded in the no-Reagan, rerun conditions of history.[8]
d. How confident can we be—in the spring of 2004—that the Bush administration was myopic to the threat posed by Al Qaeda in the summer of 2001, failing to heed classified memos that baldly announced "bin Laden plans to attack the United States"? Or is all this 20/20 hindsight motivated by desire to topple a president? Have we forgotten how vague the warnings were, how vocal the outcry would have been against FBI-CIA coordination, and how stunned Democrats and Republicans alike were by the attack?[9]

Where then does this leave us? Up to a disconcertingly difficult to identify point, the relativists are right: judgments of political judgment can never be rendered politically uncontroversial. Many decades of case study experience should by now have drummed in the lesson that one observer's simpleton will often be another's man of principle; one observer's groupthink, another's well-run meeting.

But the relativist critique should not paralyze us. It would be a massive mistake to "give up," to approach good judgment solely from first-person pronoun perspectives that treat our own intuitions about what constitutes good judgment, and about how well we stack up against those intuitions, as the beginning and end points of inquiry.

This book is predicated on the assumption that, even if we cannot capture all of the subtle counterfactual and moral facets of good judgment, we can advance the cause of holding political observers accountable to independent standards of empirical accuracy and logical rigor. Whatever their allegiances, good judges should pass two types of tests:

[8] J. Matlock, *Autopsy on an Empire: the American Ambassador's Account of the Collapse of the Soviet Union* (New York: Random House, 1995); B. Farnham, "Perceiving the End of Threat: Ronald Reagan and the Gorbachev Revolution," in *Good Judgment in Foreign Policy*, 153–90. R. L. Garthoff, *The Great Transition: American-Soviet Relations and the End of the Cold War* (Washington, DC: Brookings Institution, 1994).

[9] The debate on this case has only begun. But the 9/11 Presidential Commission has laid out a thoughtful framework for conducting it (The 9/11 Commission Report. New York: Norton, 2004).

1. Correspondence tests rooted in empiricism. How well do their private beliefs map onto the publicly observable world?
2. Coherence and process tests rooted in logic. Are their beliefs internally consistent? And do they update those beliefs in response to evidence?

In plain language, good judges should both "get it right" and "think the right way."[10]

This book is also predicated on the assumption that, to succeed in this ambitious undertaking, we cannot afford to be parochial. Our salvation lies in multimethod triangulation—the strategy of pinning down elusive constructs by capitalizing on the complementary strengths of the full range of methods in the social science tool kit. Our confidence in specific claims should rise with the quality of converging evidence we can marshal from diverse sources. And, insofar as we advance many interdependent claims, our confidence in the overall architecture of our argument should be linked to the sturdiness of the interlocking patterns of converging evidence.[11]

Of course, researchers are more proficient with some tools than others. As a research psychologist, my comparative advantage does not lie in doing case studies that presuppose deep knowledge into the challenges confronting key players at particular times and places.[12] It lies in applying the distinctive skills that psychologists collectively bring to this challenging topic: skills honed by a century of experience in translating vague speculation about human judgment into testable propositions. Each chapter of

[10] On the fundamental status of correspondence and coherence standards in judging judgment, see K. Hammond, *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice* (New York: Oxford University Press, 1996).

[11] This project offers many examples of interlocking convergence: our hedgehog-fox measure of cognitive style predicts indicators of good judgment similar to those predicted by kindred measures elsewhere; our qualitative analysis of forecasters' explanations for their predictions dovetails with our quantitative analyses of why foxes outperformed hedgehogs; our findings of poky belief updating among forecasters, especially hedgehogs, mesh well with laboratory research on "cognitive conservatism." Psychologists will see here the cumulative logic of construct validation. See D. T. Campbell and D. W. Fiske, "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin* 56 (1959): 81–105.

[12] I avoid ambitious conceptions of good judgment that require, for instance, my judging how skillfully policy makers juggle trade-offs among decision quality (is this policy the best policy given our conception of national interest?), acceptability (can we sell this policy?), and timeliness (how should we factor in the costs of delay?). (A. L. George, *Presidential Decision-Making in Foreign Policy* [Boulder, CO: Westview, 1980]) I also steer clear of conceptions that require my judging whether decision makers grasped "the essential elements of a problem and their significance" or "considered the full range of viable options." (S. Renshon, "Psychological Sources of Good Judgment in Political Leaders, in *Good Judgment in Foreign Policy*, 25–57).

this book exploits concepts from experimental psychology to infuse the abstract goal of assessing good judgment with operational substance, so we can move beyond anecdotes and calibrate the accuracy of observers' predictions, the soundness of the inferences they draw when those predictions are or are not borne out, the evenhandedness with which they evaluate evidence, and the consistency of their answers to queries about what could have been or might yet be.[13]

The goal was to discover how far back we could push the "doubting Thomases" of relativism by asking large numbers of experts large numbers of questions about large numbers of cases and by applying no-favoritism scoring rules to their answers. We knew we could never fully escape the interpretive controversies that flourish at the case study level. But we counted on the law of large numbers to cancel out the idiosyncratic case-specific causes for forecasting glitches and to reveal the invariant properties of good judgment.[14] The miracle of aggregation would give us license to tune out the kvetching of sore losers who, we expected, would try to justify their answers by arguing that our standardized questions failed to capture the subtleties of particular situations or that our standardized scoring rules failed to give due credit to forecasts that appear wrong to the uninitiated but that are in some deeper sense right.

The results must speak for themselves, but we made progress down this straight and narrow positivist path. We can construct multimethod composite portraits of good judgment in chapters 3, 4, and 5 that give zero weight to complaints about the one-size-fits-all ground rules of the project and that pass demanding statistical tests. If I had stuck to this path, my life would have been simpler, and this book shorter. But, as I listened to the counterarguments advanced by the thoughtful professionals who participated in this project, it felt increasingly high-handed to dismiss every complaint as a squirmy effort to escape disconfirmation.

[13] My approach represents a sharp shift away from case-specific "idiographic" knowledge (who gets what right at specific times and places?) toward more generalizable or "nomothetic" knowledge (who tends to be right across times and places?). Readers hoping for the scoop on who was right about "shock therapy" or the "Mexican bailout" will be disappointed. Readers should stay tuned, though, if they are curious why some observers manage to assign consistently more realistic probabilities across topics.

[14] The law of large numbers is a foundational principle of statistics, and Stigler traces it to the eighteenth century. He quotes Bernoulli: "For even the most stupid of men, by some instinct of nature . . . is convinced that the more observations have been made, the less danger there is of wandering from one's goal." And Poisson: "All manner of things are subject to a universal law that we may call the law of large numbers . . . : if we observe a large number of events of the same nature, dependent upon constant causes and upon causes that vary irregularly . . . we will find the ratios between the numbers of these events are approximately constant." (S. Stigler, 1986, *The History of Statistics: The Measurement of Uncertainty Before 1900* [Cambridge: Harvard University Press, 1986], 65, 185)

My participants knew my measures—however quantitative the veneer—were fallible. They did not need my permission to argue that the flaws lay in my procedures, not in their answers.

We confronted more and more judgment calls on how far to go in accommodating these protests. And we explored more and more adjustments to procedures for scoring the accuracy of experts' forecasts, including *value adjustments* that responded to forecasters' protests that their mistakes were the "right mistakes" given the costs of erring in the other direction; *controversy adjustments* that responded to forecasters' protests that they were really right and our reality checks wrong; *difficulty adjustments* that responded to protests that some forecasters had been dealt tougher tasks than others; and even *fuzzy-set adjustments* that gave forecasters partial credit whenever they claimed that things that did not happen either almost happened or might yet happen.

We could view these scoring adjustments as the revenge of the relativists. The list certainly stretches our tolerance for uncertainty: it requires conceding that the line between rationality and rationalization will *often* be blurry. But, again, we should not concede too much. Failing to learn everything is not tantamount to learning nothing. It is far more reasonable to view the list as an object lesson in how science works: tell us your concerns and we will translate them into scoring procedures and estimate how sensitive our conclusions about good judgment are to various adjustments. Indeed, these sensitivity analyses will reveal the composite statistical portraits of good judgment to be robust across an impressive range of scoring adjustments, with the conditional likelihood of such patterns emerging by chance well under five in one hundred (likelihood conditional on null hypothesis being true).

No number of statistical tests will, however, compel principled relativists to change their minds about the propriety of holding advocates of clashing worldviews accountable to common standards—a point we drive home in the stock-taking closing chapter. But, in the end, most readers will not be philosophers—and fewer still relativists.

This book addresses a host of more pragmatic audiences who have learned to live with the messy imperfections of social science (and be grateful when the epistemological glass is one-third full rather than annoyed about its being two-thirds empty). Our findings will speak to psychologists who wonder how well laboratory findings on cognitive styles, biases, and correctives travel in the real world, decision theorists who care about the criteria we use for judging judgment, political scientists who wonder who has what it takes to "bridge the gap" between academic abstractions and the real world, and journalists, risk consultants, and intelligence analysts who make their livings thinking in "real time" and might be curious who can "beat" the dart-throwing chimp.

I can promise these audiences tangible "deliverables." We shall learn how to design correspondence and coherence tests that hold pundits more accountable for their predictions, even if we cannot whittle their wiggle room down to zero. We shall learn why "what experts think" is so sporadic a predictor of forecasting accuracy, why "how experts think" is so consistent a predictor, and why self-styled foxes outperformed hedgehogs on so wide a range of tasks, with one key exception where hedgehogs seized the advantage. Finally, we shall learn how this patterning of individual differences sheds light on a fundamental trade-off in all historical reasoning: the tension between defending our worldviews and adapting those views to dissonant evidence.

Tracking Down an Elusive Construct

Announcing bold intentions is easy. But delivering is hard: it requires moving beyond vague abstractions and spelling out how one will measure the intricate correspondence and coherence facets of the multifaceted concept of good judgment.

Getting It Right

Correspondence theories of truth identify good judgment with the goodness of fit between our internal mental representations and corresponding properties of the external world. Just as our belief that grass is green owes its truth to an objective feature of the physical world—grass reflects a portion of the electromagnetic spectrum visible to our eyes—the same can be said for beliefs with less precise but no less real political referents: wars break out, economies collapse. We should therefore credit good judgment to those who see the world as it is—or soon will be.[15] Two oft-derived corollaries are: (1) we should bestow bonus credit on those farsighted souls who saw things well before the rest of us—the threat posed by Hitler in the early 1930s or the vulnerability of the Soviet Union in the early 1980s or the terrorist capabilities of radical Islamic organizations in the 1990s or the puncturing of the Internet bubble in 2000; (2) we should penalize those misguided souls who failed to see

[15] Our correspondence measures focused on the future, not the present or past, because we doubted that the sophisticated specialists in our sample would make the crude partisan errors of fact ordinary citizens make (see D. Green, B. Palmquist, and E. Schickler, *Partisan Hearts and Minds* [New Haven, CT: Yale University Press, 2002]). Pilot testing confirmed these doubts. Even the most dogmatic Democrats in our sample knew that inflation fell in the Reagan years, and even the most dogmatic Republicans knew that budget deficits shrank in the Clinton years. To capture susceptibility to biases among our respondents, we needed a more sophisticated mousetrap.

things long after they became obvious to the rest of us—who continued to believe in a monolithic Communist bloc long after the Sino-Soviet rupture or in Soviet expansionism through the final Gorbachev days.

Assessing this superficially straightforward conception of good judgment proved, however, a nontrivial task. We had to pass through a gauntlet of five challenges.[16]

1. *Challenging whether the playing fields are level.* We risk making false attributions of good judgment if some forecasters have been dealt easier tasks than others. Any fool can achieve close to 100 percent accuracy when predicting either rare outcomes, such as nuclear proliferation or financial collapse, or common ones, such as regular elections in well-established democracies. All one need do is constantly predict the higher base rate outcome and—like the proverbial broken clock—one will look good, at least until skeptics start benchmarking one's performance against simple statistical algorithms.
2. *Challenging whether forecasters' "hits" have been purchased at a steep price in "false alarms."* We risk making false attributions of good judgment if we fixate solely on success stories—crediting forecasters for spectacular hits (say, predicting the collapse of the Soviet Union) but not debiting them for false alarms (predicting the disintegration of nation-states—e.g., Nigeria, Canada—still with us). Any fool can also achieve high hit rates for any outcome—no matter how rare or common—by indiscriminately attaching high likelihoods to its occurrence. We need measures that take into account all logically possible prediction-outcome matchups: saying x when x happens (hit); saying x when x fails to happen (false alarm or overprediction); saying $\sim x$ when $\sim x$ happens (correct rejection); and saying $\sim x$ when x happens (miss or underprediction).
3. *Challenging the equal weighting of hits and false alarms.* We risk making false attributions of good judgment if we treat political reasoning as a passionless exercise of maximizing aggregate accuracy. It is profoundly misleading to talk about forecasting accuracy without spelling out the trade-offs that forecasters routinely make between the conflicting risks of overprediction (false alarms: assigning high probabilities to events that do not occur) and underprediction (misses: assigning low probabilities to events that do occur).[17] Consider but two illustrations:

[16] For thoughtful discussions of correspondence measures, see A. Kruglanski, *Lay Epistemics and Human Knowledge* (New York: Plenum Press, 1989; D. A. Kenny, *Interpersonal Perception* (New York: Guilford Press, 1994).

[17] John Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics* (Mahwah, NJ: Lawrence Erlbaum, 1996).

a. Conservatives in the 1980s justified their suspicions of Gorbachev by insisting that underestimating Soviet strength was the more serious error, tempting us to relax our guard and tempting them to test our resolve. By contrast, liberals worried that overestimating the Soviets would lead to our wasting vast sums on superfluous defense programs and to our reinforcing the Soviets' worst-case suspicions about us.
b. Critics of the Western failure to stop mass killings of the 1990s in Eastern Europe or central Africa have argued that, if politicians abhorred genocide as much as they profess in their brave "never again" rhetoric, they would have been more sensitive to the warning signs of genocide than they were. Defenders of Western policy have countered that the cost of false-alarm intrusions into the internal affairs of sovereign states would be prohibitive, sucking us into a succession of Vietnam-style quagmires.

Correspondence indicators are, of course, supposed to be value neutral, to play no favorites and treat all mistakes equally. But we would be remiss to ignore the possibility we are misclassifying as "wrong" forecasters who have made value-driven decisions to exaggerate certain possibilities. Building on past efforts to design correspondence indicators that are sensitive to trade-offs that forecasters strike between over- and underprediction, the Technical Appendix lays out an array of value adjustments that give forecasters varying benefits of the doubt that their mistakes were the "right mistakes."[18]

4. *Challenges of scoring subjective probability forecasts.* We cannot assess the accuracy of experts' predictions if we cannot figure out what they predicted. And experts were reluctant to call outcomes either impossible or inevitable. They hedged with expressions such as "remote chance," "maybe," and "odds-on favorite." Checking the correctness of vague verbiage is problematic. Words can take on many meanings: "likely" could imply anything from barely better than 50/50 to 99 percent.[19] Moreover, checking the correctness of numerical probability estimates is problematic. Only judgments of zero (impossible) and 1.0 (inevitable) are technically falsifiable. For all other values, wayward forecasters can argue that we stumbled into improbable worlds: low-probability events sometimes happen and high-probability events sometimes do not.

   To break this impasse, we turned to behavioral decision theorists who have had success in persuading other reluctant professionals to translate verbal waffling into numerical probabilities as well as in scoring these judgments.[20] The key insight is that, although we can never know whether there was a .1 chance in 1988 that the Soviet Union would disintegrate by 1993 or a .9 chance of Canada disintegrating by 1998, we can measure the accuracy of such judgments across many events (saved again by the law of large numbers). These aggregate measures tell us how discriminating forecasters were: do they assign larger probabilities to things that subsequently happen than to things that do not? These measures also tell us how well calibrated forecasters were: do events they assign .10 or .50 or .90 probabilities materialize roughly 10 percent or 50 percent or 90 percent of the time? And the Technical Appendix shows us how to tweak these measures to tap into a variety of other finer-grained conceptions of accuracy.
5. *Challenging reality.* We risk making false attributions of good judgment if we fail to recognize the existence of legitimate ambiguity about either what happened or the implications of what happened for the truth or falsity of particular points of view.

Perfect consensus over what happened is often beyond reach. Partisan Democrats and Republicans will remain forever convinced that the pithiest characterization of the 2000 presidential election is that the other side connived with judicial hacks to steal it. Rough agreement is, however, possible as long as we specify outcomes precisely enough to pass the litmus tests in the Methodological Appendix. The most important of these was the clairvoyance test: our measures had to define possible futures so clearly that, if we handed experts' predictions to a true clairvoyant, she could tell us, with no need for clarifications ("What did you

[18] J. Swets, R. Dawes, and J. Monahan, "Psychological Science Can Improve Diagnostic Decisions, *Psychological Science in the Public Interest*, 1 (2000): 1–26. These mental exercises compel us to be uncomfortably explicit about our priorities. Should we give into the utilitarian temptation to save lives by ending a long war quickly via a tactical nuclear strike to "take out" the enemy leadership? Or should we define good judgment as the refusal to countenance taboo trade-offs, as the wise recognition that some things are best left unthinkable? See P. E. Tetlock, O. Kristel, B. Elson, M. Green, and J. Lerner, (2000). "The Psychology of the Unthinkable: Taboo Trade-Offs, Forbidden Base Rates, and Heretical Counterfactuals, *Journal of Personality and Social Psychology*, 78 (2000): 853–70.

[19] Many studies have examined the varied meanings that people attach to verbal expressions of uncertainty: W. Bruine de Bruin, B. Fischhoff, S. G. Millstein, and B. L. Felscher, "Verbal and Numerical Expressions of Probability: 'It's a Fifty-Fifty Chance.'" *Organizational Behavior and Human Decision Processes* 81 (2000): 115–23.

[20] The pioneering work focused on weather forecasters. See A. H. Murphy, "Scalar and Vector Partitions of the Probability Score, Part I, Two-Stage Situation," *Journal of Applied Meteorology* 11 (1972): 273–82; A. H. Murphy, "Scalar and Vector Partitions of the Probability Score, Part II, N-State Situation," *Journal of Applied Meteorology* 12 (1972): 595–600. For extensions, see R. L. Winkler, "Evaluating Probabilities: Asymmetric Scoring Rules," *Management Science* 40 (1994): 1395–1405.

mean by a Polish Peron or . . . ?"), who got what right. This test rules out oracular pronouncements of the Huntington or Fukuyama sort: expect clashes of civilizations or end of history. Our measures were supposed to focus, to the degree possible,[21] on the unadorned facts, the facts before the spinmeisters dress them up: before "defense spending as percentage of GDP" is rhetorically transformed into "reckless warmongering" or "prudent precaution."

The deeper problem—for which there is no ready measurement fix—is resolving disagreements over the implications of what happened for the correctness of competing points of view. Well before forecasters had a chance to get anything wrong, many warned that forecasting was an unfair standard—unfair because of the danger of lavishing credit on winners who were just lucky and heaping blame on losers who were just unlucky.

These protests are not just another self-serving effort of ivory tower types to weasel out of accountability to real-world evidence. Prediction and explanation are not as tightly coupled as once supposed.[22] Explanation is possible without prediction. A conceptually trivial but practically consequential source of forecasting failure occurs whenever we possess a sound theory but do not know whether the antecedent conditions for applying the theory have been satisfied: high school physics tells me why the radiator will freeze if the temperature falls below 32°F but not how cold it will be tonight. Or, consider cases in which we possess both sound knowledge and good knowledge of antecedents but are stymied because outcomes may be subject to chaotic oscillations. Geophysicists understand how principles of plate tectonics produce earthquakes and can monitor seismological antecedents but still cannot predict earthquakes.

Conversely, prediction is possible without explanation. Ancient astronomers had bizarre ideas about what stars were, but that did not stop them from identifying celestial regularities that navigators used to guide ships for centuries. And contemporary astronomers can predict the rhythms of solar storms but have only a crude understanding of what causes these potentially earth-sizzling eruptions. For most scientists, prediction is not enough. Few scientists would have changed their minds about astrology if Nancy Reagan's astrologer had chalked up a string of spectacular forecasting successes. The result so undercuts core beliefs that the scientific community would have, rightly, insisted on looking long and hard for other mechanisms underlying these successes.

[21] The caveat is critical. The more experts knew, the harder it often became to find indicators that passed the clairvoyance test. For instance, GDP can be estimated in many ways (we rely on purchasing power parity), and so can defense spending.

[22] F. Suppe, *The Structure of Scientific Theories* (Chicago: University of Chicago Press, 1973); S. Toulmin, *Foresight and Understanding: An Inquiry into the Aims of Science* (New York: Harper & Row, 1963).

These arguments highlight valid objections to simple correspondence theories of truth. And the resulting complications create far-from-hypothetical opportunities for mischief. It is no coincidence that the explanation-is-possible-without-prediction argument surges in popularity when our heroes have egg on their faces. Pacifists do not abandon Mahatma Gandhi's worldview just because of the sublime naïveté of his remark in 1940 that he did not consider Adolf Hitler to be as bad as "frequently depicted" and that "he seems to be gaining his victories without much bloodshed";[23] many environmentalists defend Paul Ehrlich despite his notoriously bad track record in the 1970s and 1980s (he predicted massive food shortages just as new technologies were producing substantial surpluses);[24] Republicans do not change their views about the economic competence of Democratic administrations just because Martin Feldstein predicted that the legacy of the Clinton 1993 budget would be stagnation for the rest of the decade;[25] social democrats do not overhaul their outlook just because Lester Thurow predicted that the 1990s would witness the ascendancy of the more compassionate capitalism of Europe and Japan over the "devil take the hindmost" American model.[26]

Conversely, it is no coincidence that the prediction-is-possible-without-explanation argument catches on when our adversaries are crowing over their forecasting triumphs. Our adversaries must have been as lucky in victory as we were unlucky in defeat. After each side has taken its pummeling in the forecasting arena, it is small wonder there are so few fans of forecasting accuracy as a benchmark of good judgment.

Such logical contortions should not, however, let experts off the hook. Scientists ridicule explanations that redescribe past regularities as empty tautologies—and they have little patience with excuses for consistently poor predictive track records. A balanced assessment would recognize that forecasting is a fallible but far from useless indicator of our understanding of causal mechanisms. In the long run (and we solicit enough forecasts on enough topics that the law of large numbers applies), our confidence in a point of view should wax or wane with its predictive successes and failures, the exact amounts hinging on the aggressiveness of forecasters' ex ante theoretical wagers and on our willingness to give weight to forecasters' ex post explanations for unexpected results.

[23] C. Cerf, and V. S. Navasky, eds., *The Experts Speak: The Definitive Compendium of Authoritative Misinformation* (New York: Pantheon Books, 1984).

[24] A. Sen, *Poverty and Famines* (New York: Oxford University Printing House, 1981).

[25] M. Feldstein, "Clinton's Revenue Mirage," *Wall Street Journal*, April 6, 1993, A14.

[26] See Lester Thurow, *Head to Head: The Coming Economic Battle among Japan, Europe, and America* (New York: Murrow, 1992).

Thinking the Right Way

One might suppose there must be close ties between correspondence and coherence/process indicators of good judgment, between getting it right and thinking the right way. There are connections but they are far from reliably deterministic. One could be a poor forecaster who works within a perfectly consistent belief system that is utterly detached from reality (e.g., paranoia). And one could be an excellent forecaster who relies on highly intuitive but logically indefensible guesswork.

One might also suppose that, even if our best efforts to assess correspondence indicators bog down in disputes over what really or nearly happened, we are on firmer ground with coherence/process indicators. One would again be wrong. Although purely logical indicators command deference, we encounter resistance even here. It is useful to array coherence/process indicators along a rough controversy continuum anchored at one end by widely accepted tests and at the other by bitterly contested ones.

At the close-to-slam-dunk end, we find violations of logical consistency so flagrant that few rise to their defense. The prototypic tests involve breaches of axiomatic identities within probability theory.[27] For instance, it is hard to defend forecasters who claim that the likelihood of a set of outcomes, judged as a whole, is less than the sum of the separately judged likelihoods of the set's exclusive and exhaustive membership list.[28] Insofar as there are disputes, they center on how harshly to judge these mistakes: whether people merely misunderstood instructions or whether the mistakes are by-products of otherwise adaptive modes of thinking or whether people are genuinely befuddled.

At the controversial end of the continuum, competing schools of thought offer unapologetically opposing views on the standards for judging judgment. These tests are too subjective for my taste, but they foreshadow later controversies over cognitive styles. For instance, the more committed observers are to parsimony, the more critical they are of those who fail to organize their belief systems in tidy syllogisms that deduce historical outcomes from covering laws and who flirt with close-call counterfactuals that undercut basic "laws of history"; conversely, the less committed observers are to parsimony, the more critical they are of the "rigidity" of those who try to reduce the quirkiness of history to theoretical formulas. One side's rigor is the other's dogmatism.

[27] L. Savage, *The Foundations of Statistics* (New York: Wiley, 1954); W. Edwards, "The Theory of Decision Making," *Psychological Bulletin* 51 (1954): 380–417.

[28] It requires little ingenuity to design bets that turn violators of this minimalist standard of rationality into money pumps. People do, however, often stumble. See A. Tversky, and D. Koehler, "Support Theory: A Nonextensional Representation of Subjective Probability," *Psychological Review* 101 (1994): 547–67.

In the middle of the continuum, we encounter consensus on what it means to fail coherence/process tests but divisions on where to locate the pass-fail cutoffs. The prototypic tests involve breaches of rules of fair play in the honoring of *reputational bets* and in the evenhanded treatment of evidence in *turnabout thought experiments*.

To qualify as a good judge within a Bayesian framework—and many students of human decision making as well as high-IQ public figures such as Bill Gates and Robert Rubin think of themselves as Bayesians—one must own up to one's reputational bets. The Technical Appendix lays out the computational details, but the core idea is a refinement of common sense. Good judges are good belief updaters who follow through on the logical implications of reputational bets that pit their favorite explanations against alternatives: if I declare that x is .2 likely if my "theory" is right and .8 likely if yours is right, and x occurs, I "owe" some belief change.[29]

In principle, no one disputes we should change our minds when we make mistakes. In practice, however, outcomes do not come stamped with labels indicating whose forecasts have been disconfirmed. Chapter 4 shows how much wiggle room experts can create for themselves by invoking various belief system defenses. Forecasters who expected the demise of Canada before 2000 can argue that Quebec almost seceded and still might. And Paul Ehrlich, a "doomster" known for his predictions of ecocatastrophes, saw no need whatsoever to change his mind after losing a bet with "boomster" Julian Simon over whether real prices of five commodities would increase in the 1980s. After writing a hefty check to Simon to cover the cost spread on the futures contracts, Ehrlich defiantly compared Simon to a man who jumps from the Empire State Building and, as he passes onlookers on the fiftieth floor, announces, "All's well so far."[30]

How should we react to such defenses? Philosophers of science who believe in playing strictly by ex ante rules maintain that forecasters who rewrite their reputational bets, ex post, are sore losers. Sloppy relativism will be the natural consequence of letting us change our minds—whenever convenient—on what counts as evidence. But epistemological liberals will demur. Where is it written, they ask, that we cannot revise reputational bets, especially in fuzzy domains where the truth is rarely either-or? A

[29] P. E. Tetlock, "Theory-Driven Reasoning about Possible Pasts and Possible Futures," *American Journal of Political Science* 43 (1999): 335–36. Sherman Kent, the paragon of intelligence analysts, was an early advocate of translating vague hunches into precise probabilistic odds (S. Kent, *Collected Essays* (U.S. Government: Center for the Study of Intelligence, 1970), http://www.cia.gov/csi/books/shermankent/toc.html.

[30] For an account of the Ehrlich-Simon bet, see John Tierney, "Betting on the Planet," *New York Times Magazine*, December 2, 1990, 52–53, 74–81.

balanced assessment here would concede that Bayesians can no more purge subjectivity from coherence assessments of good judgment than correspondence theorists can ignore complaints about the scoring rules for forecasting accuracy. But that does not mean we cannot distinguish desperate patch-up rewrites that delay the day of reckoning for bankrupt ideas from creative rewrites that stop us from abandoning good ideas.[31] Early warning signs that we are slipping into solipsism include the frequency and self-serving selectivity with which we rewrite bets and the revisionist scale of the rewrites.

Shifting from forward-in-time reasoning to backward-in-time reasoning, we relied on turnabout thought experiments to assess the willingness of analysts to change their opinions on historical counterfactuals. The core idea is, again, simple. Good judges should resist the temptation to engage in self-serving reasoning when policy stakes are high and reality constraints are weak. And temptation is ubiquitous. Underlying all judgments of whether a policy was shrewd or foolish are hidden layers of speculative judgments about how history would have unfolded had we pursued different policies.[32] We have warrant to praise a policy as great when we can think only of ways things could have worked out far worse, and warrant to call a policy disastrous when we can think only of ways things could have worked out far better. Whenever someone judges something a failure or success, a reasonable rejoinder is: "Within what distribution of possible worlds?"[33]

Turnabout thought experiments gauge the consistency of the standards that we apply to counterfactual claims. We fail turnabout tests when we apply laxer standards to evidence that reinforces as opposed to undercuts our favorite what-if scenarios. But, just as some forward-in-time reasoners balked at changing their minds when they lost reputational bets, some backward-in-time reasoners balked at basing their assessments of the probative value of archival evidence solely on information available before they knew how the evidence would break. They argued that far-fetched claims require stronger evidence than claims they felt had strong support from other sources. A balanced assessment here requires confronting a dilemma: if we only accept evidence that confirms

[31] Suppe, *The Structure of Scientific Theories*; P. Laudan, *Progress and Its Problems* (Berkeley: University of California Press, 1986).

[32] We discover how reliant we are on hidden counterfactuals when we probe the underpinnings of attributions of good or bad judgment to leaders. The simplest rule—"If it happens on your watch . . ."—has the advantage of reducing reliance on counterfactuals but the disadvantage of holding policy makers accountable for outcomes outside their control. Most of us want leeway for the possibilities that (a) some leaders do all the right things but—by bad luck—get clobbered; (b) other leaders violate all the rules of rationality and—by sheer dumb luck—prosper.

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

our worldview, we will become prisoners of our preconceptions, but if we subject all evidence, agreeable or disagreeable, to the same scrutiny, we will be overwhelmed. As with reputational bets, the question becomes how much special treatment of favorite hypotheses is too much. And, as with reputational bets, the bigger the double standard, the greater are the grounds for concern.

Preview of Chapters to Follow

The bulk of this book is devoted to determining how well experts perform against this assortment of correspondence and coherence benchmarks of good judgment.

Chapters 2 and 3 explore correspondence indicators. Drawing on the literature on judgmental accuracy, I divide the guiding hypotheses into two categories: those rooted in *radical skepticism*, which equates good political judgment with good luck, and those rooted in *meliorism*, which maintains that the quest for predictors of good judgment, and ways to improve ourselves, is not quixotic and there are better and worse ways of thinking that translate into better and worse judgments.

Chapter 2 introduces us to the radical skeptics and their varied reasons for embracing their counterintuitive creed. Their guiding precept is that, although we often talk ourselves into believing we live in a predictable world, we delude ourselves: history is ultimately one damned thing after another, a random walk with upward and downward blips but devoid of thematic continuity. Politics is no more predictable than other games of chance. On any given spin of the roulette wheel of history, crackpots will claim vindication for superstitious schemes that posit patterns in randomness. But these schemes will fail in cross-validation. What works today will disappoint tomorrow.[34]

Here is a doctrine that runs against the grain of human nature, our shared need to believe that we live in a comprehensible world that we can master if we apply ourselves.[35] Undiluted radical skepticism requires us to believe, *really believe*, that when the time comes to choose among

[34] The exact time of arrival of disappointment may, though, vary. The probability of black or red on a roulette spin should be independent of earlier spins. But political-economic outcomes are often interdependent. If one erroneously predicted the rise of a "Polish Peron," one would have also been wrong about surging central government debt-to-GDP ratios, inflation, corruption ratings, and so on. Skeptics should predict as much consistency in who gets what right as there is interdependence among outcomes.

[35] Radical skepticism as defined here should not be confused with radical relativism as defined earlier. Radical skeptics do not doubt the desirability or feasibility of holding different points of view accountable to common correspondence and coherence tests; they doubt only that, when put to these tests, experts can justify their claims to expertise.

controversial policy options—to support Chinese entry into the World Trade Organization or to bomb Baghdad or Belgrade or to build a ballistic missile defense—we could do as well by tossing coins as by consulting experts.[36]

Chapter 2 presents evidence from regional forecasting exercises consistent with this debunking perspective. It tracks the accuracy of hundreds of experts for dozens of countries on topics as disparate as transitions to democracy and capitalism, economic growth, interstate violence, and nuclear proliferation. When we pit experts against minimalist performance benchmarks—dilettantes, dart-throwing chimps, and assorted extrapolation algorithms—we find few signs that expertise translates into greater ability to make either "well-calibrated" or "discriminating" forecasts.

Radical skeptics welcomed these results, but they start squirming when we start finding patterns of consistency in who got what right. Radical skepticism tells us to expect nothing (with the caveat that if we toss enough coins, expect some streakiness). But the data revealed more consistency in forecasters' track records than could be ascribed to chance. Meliorists seize on these findings to argue that crude human-versus-chimp comparisons mask systematic individual differences in good judgment.

Although meliorists agree that skeptics go too far in portraying good judgment as illusory, they agree on little else. Cognitive-content meliorists identify good judgment with a particular outlook but squabble over which points of view represent movement toward or away from the truth. Cognitive-style meliorists identify good judgment not with *what* one thinks, but with *how* one thinks. But they squabble over which styles of reasoning—quick and decisive versus balanced and thoughtful—enhance or degrade judgment.

Chapter 3 tests a multitude of meliorist hypotheses—most of which bite the dust. *Who* experts were—professional background, status, and so on—made scarcely an iota of difference to accuracy. Nor did *what* experts thought—whether they were liberals or conservatives, realists or institutionalists, optimists or pessimists. But the search bore fruit. *How* experts thought—their style of reasoning—did matter. Chapter 3 demonstrates the usefulness of classifying experts along a rough cognitive-style continuum anchored at one end by Isaiah Berlin's prototypical hedgehog and at the other by his prototypical fox.[37] The intellectually aggressive hedgehogs knew one big thing and sought, under the banner of parsimony,

[36] The unpalatability of a proposition is weak grounds for rejecting it. But it often influences where we set our thresholds of proof. (P. E. Tetlock, "Political or Politicized Psychology: Is the Road to Scientific Hell Paved with Good Moral Intentions?" *Political Psychology* 15 [1994]: 509–30)

[37] Berlin, "The Hedgehog and the Fox."

to expand the explanatory power of that big thing to "cover" new cases; the more eclectic foxes knew many little things and were content to improvise ad hoc solutions to keep pace with a rapidly changing world.

Treating the regional forecasting studies as a decathlon between rival strategies of making sense of the world, the foxes consistently edge out the hedgehogs but enjoy their most decisive victories in long-term exercises inside their domains of expertise. Analysis of explanations for their predictions sheds light on how foxes pulled off this cognitive-stylistic coup. The foxes' self-critical, point-counterpoint style of thinking prevented them from building up the sorts of excessive enthusiasm for their predictions that hedgehogs, especially well-informed ones, displayed for theirs. Foxes were more sensitive to how contradictory forces can yield stable equilibria and, as a result, "overpredicted" fewer departures, good or bad, from the status quo. But foxes did not mindlessly predict the past. They recognized the precariousness of many equilibria and hedged their bets by rarely ruling out anything as "impossible."

These results favor meliorism over skepticism—and they favor the pro-complexity branch of meliorism, which proclaims the adaptive superiority of the tentative, balanced modes of thinking favored by foxes,[38] over the pro-simplicity branch, which proclaims the superiority of the confident, decisive modes of thinking favored by hedgehogs.[39] These results also domesticate radical skepticism, with its wild-eyed implication that experts have nothing useful to tell us about the future beyond what we could have learned from tossing coins or inspecting goat entrails. This tamer brand of skepticism—skeptical meliorism—still warns of the dangers of hubris, but it allows for how a self-critical, dialectical style of reasoning can spare experts the big mistakes that hammer down the accuracy of their more intellectually exuberant colleagues.

Chapter 4 shifts the spotlight from whether forecasters get it right to whether forecasters change their minds as much as they should when they get it wrong. Using experts' own reputational bets as our benchmark, we discover that experts, especially the hedgehogs, were slower than they should have been in revising the guiding ideas behind inaccurate forecasts.[40] Chapter 4 also documents the belief system defenses that experts use to justify rewriting their reputational bets after the fact: arguing that, although the predicted event did not occur, it eventually will

[38] For a review of work on cognitive styles, see P. Suedfeld, and P. E. Tetlock, "Cognitive styles," in *Blackwell International Handbook of Social Psychology: Intra-Individual Processes*, vol. 1, ed. A. Tesser and N. Schwartz (London: Blackwell, 2000).

[39] G. Gigerenzer and P. M. Todd, *Simple Heuristics That Make Us Smart* (New York: Oxford University Press, 2000).

[40] H. J. Einhorn and R. M. Hogarth, "Prediction, Diagnosis and Causal Thinking in Forecasting," *Journal of Forecasting* 1 (1982): 23–36.

(off on timing) or it nearly did (the close call) and would have but for . . . (the exogenous shock). Bad luck proved a vastly more popular explanation for forecasting failure than good luck proved for forecasting success.

Chapter 5 lengthens the indictment: hedgehogs are more likely than foxes to uphold double standards for judging historical counterfactuals. And this double standard indictment is itself double-edged. First, there is the selective openness toward close-call claims. Whereas chapter 4 shows that hedgehogs only opened to close-call arguments that insulated their forecasts from disconfirmation (the "I was almost right" defense), chapter 5 shows that hedgehogs spurn similar indeterminacy arguments that undercut their favorite lessons from history (the "I was not almost wrong" defense). Second, chapter 5 shows that hedgehogs are less likely than foxes to apologize for failing turnabout tests, for applying tougher standards to agreeable than to disagreeable evidence. Their defiant attitude was "I win if the evidence breaks in my direction" but "if the evidence breaks the other way, the methodology must be suspect."

Chapters 4 and 5 reinforce a morality-tale reading of the evidence, with sharply etched good guys (the spry foxes) and bad guys (the self-assured hedgehogs). Chapter 6 calls on us to hear out the defense before reaching a final verdict. The defense raises logical objections to the factual, moral, and metaphysical assumptions underlying claims that "one group makes more accurate judgments than another" and demands difficulty, value, controversy and fuzzy-set scoring-rule adjustments as compensation. The defense also raises the psychological objection that there is no single, best cognitive style across situations.[41] Overconfidence may be essential for achieving the forecasting coups that posterity hails as visionary. The bold but often wrong forecasts of hedgehogs may be as forgivable as high strikeout rates among home-run hitters, the product of a reasonable trade-off, not grounds for getting kicked off the team. Both sets of defenses create pockets of reasonable doubt but, in the end, neither can exonerate hedgehogs of all their transgressions. Hedgehogs just made too many mistakes spread across too many topics.

Whereas chapter 6 highlighted some benefits of the "closed-minded" hedgehog approach to the world, chapter 7 dwells on some surprising

[41] For expansions of this argument, see P. E. Tetlock, R. S. Peterson, and J. M. Berry, Flattering and Unflattering Personality Portraits of Integratively Simple and Complex Managers," *Journal of Personality and Social Psychology* 64 (1993): 500–511; P. E. Tetlock and A. Tyler, "Winston Churchill's Cognitive and Rhetorical Style," *Political Psychology* 17 (1996): 149–70. P. E. Tetlock, D. Armor, and R. Peterson, "The Slavery Debate in Antebellum America: Cognitive Style, Value Conflict, and the Limits of Compromise," *Journal of Personality and Social Psychology* 66 (1994): 115–26.

costs of the "open-minded" fox approach. Consultants in the business and political worlds often use scenario exercises to encourage decision makers to let down their guards and imagine a broader array of possibilities than they normally would.[42] On the plus side, these exercises can check some forms of overconfidence, no mean achievement. On the minus side, these exercises can stimulate experts—once they start unpacking possible worlds—to assign too much likelihood to too many scenarios.[43] There is nothing admirably open-minded about agreeing that the probability of event A is less than the compound probability of A and B, or that x is inevitable but alternatives to x remain possible. Trendy open-mindedness looks like old-fashioned confusion. And the open-minded foxes are more vulnerable to this confusion than the closed-minded hedgehogs.

We are left, then, with a murkier tale. The dominant danger remains hubris, the mostly hedgehog vice of closed-mindedness, of dismissing dissonant possibilities too quickly. But there is also the danger of cognitive chaos, the mostly fox vice of excessive open-mindedness, of seeing too much merit in too many stories. Good judgment now becomes a metacognitive skill—akin to "the art of self-overhearing."[44] Good judges need to eavesdrop on the mental conversations they have with themselves as they decide how to decide, and determine whether they approve of the trade-offs they are striking in the classic exploitation-exploration balancing act, that between exploiting existing knowledge and exploring new possibilities.

Chapter 8 reflects on the broader implications of this project. From a philosophy of science perspective, there is value in assessing how far an exercise of this sort can be taken. We failed to purge all subjectivity from judgments of good judgment, but we advanced the cause of "objectification" by developing valid correspondence and coherence measures of good judgment, by discovering links between how observers think and how they fare on these measures, and by determining the robustness of these links across scoring adjustments. From a policy perspective, there is value in using publicly verifiable correspondence and coherence benchmarks to gauge the quality of public debates. The more people know about pundits' track records, the stronger the pundits' incentives to compete by improving the epistemic (truth) value of their products, not just by pandering to communities of co-believers.

[42] Peter Schwarz, *The Art of the Long View* (New York: Doubleday, 1991).

[43] For a mathematical model for understanding the effects of "unpacking" on probability judgments, A. Tversky and D. Koehler, "Support Theory: A Nonextensional Representation of Subjective Probability," *Psychological Review* 101 (1994): 547–67.

[44] H. Bloom, *Shakespeare: The Invention of the Human* (New York: Riverhead, 1998).

These are my principal arguments. Like any author, I hope they stand the test of time. I would not, however, view this project as a failure if hedgehogs swept every forecasting competition in the early twenty-first century. Indeed, this book gives reasons for expecting occasional reversals of this sort. This book will count as a failure, as a dead end, only if it fails to inspire follow-ups by those convinced they can do better.

CHAPTER 2

The Ego-deflating Challenge of Radical Skepticism

> Among all forms of mistake, prophecy is the most gratuitous.
> —GEORGE ELIOT, *Middlemarch*

> What I've said that turned out to be right will be considered obvious, and what was wrong will be humorous.
> —BILL GATES, *The Road Ahead*

IT IS COMMONPLACE to lament the sad state of political forecasting. Moreover, suspicions that the entire enterprise is intellectually bankrupt have only been fortified by the most recent forecasting fiasco: the unanimous declaration by quantitative modelers of presidential elections at the American Political Science Association in August 2000 that we could ignore the frantic rhetorical posturing of the next few months. Election campaigns are tales full of sound and fury but of no significance because of the offsetting effects of each side's propaganda broadsides. The die had been cast: Gore would defeat Bush by decisive, even landslide, margins.[1]

We revisit this incident in chapter 5, so here it must suffice to caution against drawing sweeping conclusions from a single data point. The current chapter has three missions: (1) to explore why radical skeptics believe the social science quest for predictive laws to be ill-conceived; (2) to weave their arguments into a composite set of six hypotheses, the core tenets of skepticism, that tell us what to expect when a diverse array of experts tries to predict an even more diverse array of real-world events; (3) to present evidence that suggests that, although skepticism about the predictive powers of experts is warranted, the skeptics do sometimes overreach: "who gets what right" is not just a matter of blind luck.

[1] For thoughtful postmortems, see L. M. Bartels and J. Zaller, "Presidential Vote Models: A Recount," *Political Science and Politics* 34 (2001): 9–20; M. S. Lewis-Beck, and C. Tien, "Modeling the Future: Lessons from the Gore Forecast," *Political Science and Politics* 34 (2001): 21–24; C. Wlezien, "On Forecasting the Presidential Vote," *Political Science and Politics* 34 (2001): 25–32; J. E. Campbell, "The Referendum That Didn't Happen: The Forecasts of the 2000 Presidential Election," *Political Science and Politics* 34 (2001): 33–38.