CHAPTER 12

Correspondence and Coherence: Indicators of Good Judgment in World Politics

Philip E. Tetlock
University of California, Berkeley, USA

This chapter summarizes some research results on expert political judgment that bear on debates among experimental psychologists over alleged departures from rationality in human judgment (for more details, see Tetlock & Belkin, 1996; Tetlock, 1998, 1999; Tetlock & Lebow, 2001). The results derive from a 15-year research program tracking forecasting performance, and they should generally prove heartening to those in the judgment- and decision-making community who believe that demonstrations of systematic errors and biases are not just the artifactual byproducts of laboratory trickery performed on unmotivated undergraduate conscripts. The participants in all the studies reported here were seasoned professionals who made their living by analyzing and writing about political-economic trends. We shall discover that, even when seasoned professionals are making judgments about consequential real-world events within their domains of expertise, they often fall prey to such well-known errors or biases as the following:

(1) *Overconfidence.* There is frequently a large gap between the subjective probabilities that experts assign to outcomes and the objective probabilities of those outcomes materializing (Dawes, 1998).

(2) *Cognitive conservatism.* When we compare how much experts actually change their minds in response to new evidence to how much Bayes' theorem says they should change their minds, there are numerous indications that experts are too slow to update their beliefs (Slovic, Fischhoff & Lichtenstein, 1977; Einhorn & Hogarth, 1981).

(3) *Certainty of hindsight.* Experts sometimes deny mistakes altogether. They tend to recall assigning higher subjective probabilities to those political-economic outcomes that occur than they actually assigned before learning what occurred (Fischhoff, 1975; Hawkins & Hastie, 1990).

Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making. Edited by David Hardman and Laura Macchi.

(4) *Theory-driven standards of evidence and proof.* Experts generally impose higher standards of evidence and proof on dissonant claims than they do on consonant ones (Nisbett & Ross, 1980). This double standard is particularly noticeable in the reactions that political observers have to: (a) close-call counterfactuals that imply history could easily have gone down a different path and thus have implications for the validity of theoretical, ideological or policy stances; (b) historical discoveries that bear on the plausibility of these ideologically charged close-call counterfactuals.

(5) *Systematic evidence of incoherence in subjective probability judgments.* Political observers are highly susceptible to the subadditivity effects that Tversky and Koehler's (1994) support theory predicts should be the result of decomposing sets of possible futures or possible pasts into their exclusive and exhaustive components. In violation of the extensionality principle of probability theory, people often judge the likelihood of the whole to be less, sometimes far less, than the sum of its parts.

This chapter will, however, tell more than a tale about the real-world replicability of deviations from correspondence and coherence standards of good judgment. There will be some less familiar twists and turns of the argument.

(1) We shall discover how exasperatingly difficult it is to prove beyond a reasonable doubt in the political domain that experts have erred. This is true with respect to our empirical demonstrations of both overconfidence and cognitive conservatism. Experts can—and often do—defend overconfidence in their predictions of dramatic, low-base-rate outcomes as prudent efforts to call attention to the risks of war, nuclear proliferation and economic collapse. Experts also can—and often do—defend cognitive conservatism by invoking a variety of reasons for why, in light of intervening events, they should not be bound to change beliefs to the degree specified by reputational bets (likelihood ratios) they themselves made earlier.

(2) We shall discover more evidence of systematic individual differences in susceptibility to errors and biases than is customarily uncovered in experimental research programs. Cognitive style—the strength of respondents' preferences for explanatory closure and parsimony—moderated the magnitude of several effects. Specifically, respondents who valued closure and parsimony highly were more prone to biases that were rooted in excessive faith in the predictive and explanatory power of their preconceptions—biases such as overconfidence, cognitive conservatism, certainty of hindsight and selective standards of evidence and proof.

(3) We shall discover, however, that it is a mistake to suppose that high-need-for-closure experts were at a uniform disadvantage when it came to satisfying widely upheld standards of rationality within the field of judgment and decision making. There was one major class of judgmental bias—a violation of a basic coherence standard of rationality—that our more "open-minded", low-need-for-closure respondents were more prone to exhibit: namely, the subadditivity effect linked to unpacking classes of alternative counterfactual outcomes. Respondents who did not place a high value on parsimony and explanatory closure often wound up being too imaginative and assigning too much subjective probability to too many scenarios (with the result that subjective probabilities summed to well above 1.0).

(4) Susceptibility to subadditivity effects can, as Tversky and Fox (1995) noted, render people vulnerable to exploitation by shrewder competitors, who could design bets that capitalize on the resulting logical contradictions. But there is a silver lining of sorts. The imaginative capacity to transport oneself into alternative "possible worlds" confers a measure of protection against the theory-driven biases of hindsight and retrospective determinism.

In closing, I argue for the empirical robustness of many judgmental tendencies documented in laboratory research, but I also point out the normative contestability of automatic classifications of these judgmental tendencies as errors or biases in world politics. History poses distinctive challenges to normative theorists. The political observers studied here confront poorly understood (metacognitive) trade-offs as they struggle to make sense of historical flows of hard-to-classify events that unfold only once and that have difficult-to-determine numbers of branching points. My best guess is that the price of achieving cognitive closure in quirky path-dependent systems is often rigidity, whereas the price of open-mindedness in such systems is often incoherence.

METHODOLOGICAL BACKGROUND

The methodological details of the research program on political experts are documented in Tetlock (2002). Suffice it to say here that the program has involved soliciting conditional forecasts of a wide range of political and economic outcomes since the mid-1980s, with the most sustained data collection in 1988–89 and 1992–93. Roughly 200 professionals—from academia, government and international institutions—have participated in various phases of the project. Whenever possible, I ask experts not only to make predictions within their domains of expertise, but also to venture predictions outside those domains. For example, I pressed experts on China, India or the former Soviet Union to make predictions also about Canada, South Africa and Japan, and vice versa. The resulting data provide an instructive and usually humbling baseline for assessing the predictive skill conferred by many years of professional training and expertise.

The outcomes that experts have been asked to predict have also varied widely. Prediction tasks have included the following. Will this leader or political party still be in power in 5–10 years from now? Will civil or cross-border wars break out in the next 10–25 years? Will borders change in the next 10–25 years (as a result of secession/annexation, and will change be peaceful or violent)? Will GDP per capita grow faster, slower or at the same pace in the next 3 years as in the last 3 years? What about the central-government-debt-to-GDP ratio? Inflation? Unemployment? What about fiscal spending priorities? Will defense spending increase or decrease as a percentage of central-government expenditures (relative to the last 3 years)? The entities to be predicted have included the European Monetary Union (1992–93) and over 60 countries: the former Soviet Union (1988), South Africa (1988–89), North and South Korea (1988–89), Pakistan (1992), Poland (1991), Yugoslavia (1991–92), Canada (1992–93), China (1992–93), India (1992–93), Japan (1992–93), Saudi Arabia (1992–93), Mexico (1992–93), Nigeria (1992–93), Ethiopia (1992–93), Cuba (1992–93), Brazil (1992–93) and Argentina (1992–93). Experts were not, it should be stressed, asked to make point predictions. Their assignment was to assign subjective probability estimates (0–1.0) to broad classes of possible outcomes that had been carefully selected to be exclusive and exhaustive, and pass the clairvoyance test (easy to confirm or disconfirm *ex post*).

We also did not limit data collection to judgments of possible futures. Substantial effort went into soliciting judgments of possible pasts—experts' assessments of the plausibility of counterfactual conjectures bearing on how history could or would have unfolded under various contingencies. These judgments also covered a vast range of topics. Subgroups of experts judged counterfactuals that spanned several centuries: from the early 13th century ("If the Mongols had devastated Europe as they did the civilizations of China and Islam,

the rise of European power would have been thwarted") to the mid- and late 20th century ("If Kennedy had heeded his hawkish advisers during the Cuban Missile Crisis, World War III would have been the result", or "If it were not for the pressure created by the Reagan arms buildup in the early 1980s, the USSR would still be with us today"). There is obviously no firm correspondence standard for judging the accuracy of these counterfactual conjectures, but careful analysis of these judgments does shed light on critical functional properties of expert political cognition.

BIAS NO. 1: OVERCONFIDENCE

Assessing the accuracy of subjective probability judgments of arguably unique events is problematic. If I claim that there is a 0.8 chance of Quebec's seceding from Canada between 1992 and 1997 or of the USSR's remaining intact between 1988 and 1993 or of South Africa's falling into civil war between 1989 and 1994, and those events do not materialize, I can always argue that I got the probabilities right, but the low-probability event just happened to occur. Only if I really go out on a limb, and assign the most extreme subjective probabilities on the scale, zero (x is impossible) or 1.0 (x is certain), can it be said that I have made a clearly falsifiable claim?

To get around this conundrum, calibration researchers have resorted to aggregation. It may not be possible to identify overconfidence in most individual cases, but it is possible across large numbers of forecasts. If we discover that, of all the predictions given 90 percent confidence, only 70 percent materialize, or of all those given 70 percent confidence, only 52 percent materialize, we would seem to have some warrant to claim a pattern of overconfidence. Using a number of computational procedures, including the proper quadratic scoring rule and Winkler's (1994) difficulty adjustments that control for variations across environments in the ease of predicting outcomes from simple extrapolation of base rates, we find that experts as a group tend to be overconfident. Figure 12.1 presents an illustrative calibration curve from the political-forecasting data that collapses data across 166 forecasters, 20 countries, five criterion measures and two time periods. As can be seen, some experts were more prone to overconfidence than others. The best individual-difference predictor of overconfidence was a 12-item scale that had been adapted from Arie Kruglanski's research program on the need for closure (Kruglanski & Webster, 1996) and included four additional questions that probed personal epistemologies (for example, the relative perils of overestimating or underestimating the complexity of the political world). Using a quartile split, experts with the strongest preferences for closure and parsimony were more prone than those with the weakest preferences to attach subjective likelihoods to their "most likely possible futures" that substantially exceeded the average objective likelihood of those outcomes materializing.

It should, however, be noted that, even after the fact, some experts insisted that they were justified in affixing high subjective probabilities to relatively low base-rate events—such as cross-border or civil war, border shifts triggered by secession or annexation, regime shifts triggered by coups or revolutions, and the proliferation of weapons of mass destruction. They felt justified because false alarming on "x" (saying "x" will occur when it does not) was "by far the less serious error" than missing "x" (saying "x" will not occur when it does). To paraphrase one participant whom I had thoroughly debriefed: "Several false alarms do not offset the value of being ahead of the curve in calling the disintegration of the USSR

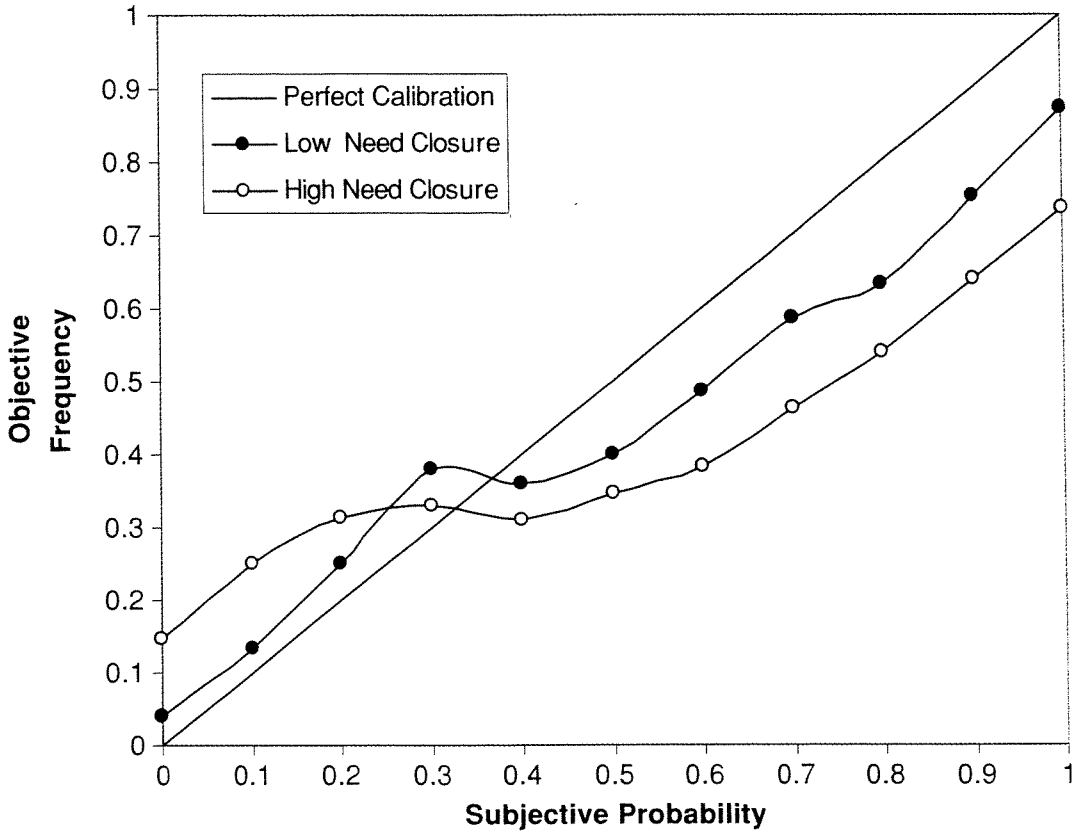

Figure 12.1 Calibration curves for experts derived by collapsing across political forecasting variables and nation-states within the 5-year forecasting frames for 1988 and 1992. Diagonal represents perfect calibration. The further a calibration curve falls below the diagonal, the greater the overconfidence

or Yugoslavia or in anticipating the nuclearization of the Indian subcontinent or the East Asian financial crisis. Who really cares if experts who are that prescient when it counts also predicted that Canada, Nigeria, Indonesia and South Africa would fall apart?" In this view, experts who hit the forecasting equivalent of home runs inevitably strike out a lot, but we should still want them on our team.

When we introduce statistical adjustments that treat "overprediction" errors as markedly less serious than "underprediction" errors for the specified outcome variables (from 1/2 to 1/4 to 1/8), the greater overconfidence effect among low-need-closure experts is significantly reduced. The effect does not, however, disappear.

BIAS NO. 2: COGNITIVE CONSERVATISM

Assessing how much experts should change their minds in response to subsequent events also raises daunting philosophical problems. Our solution took this form. For a subset of forecasting domains, experts were asked to make, *ex ante*, all the judgments necessary for constructing a reputational bet that would pit the predictive implications of their own assessments of political forces against the most influential rival perspective they cared to identify. We then plugged in Bayes' theorem to assess how much experts should have changed their minds upon learning that the outcomes either they or their

rivals deemed most likely had occurred. The standard queries for this exercise were as follows:

(1) How likely do you think each of the following sets of possible futures is if your understanding of the underlying forces at work is correct? In the Bayesian belief-updating equations, these variables go by the designations p (x1/your hypothesis), p (x2/your hypothesis)... where x1, x2... refer to sets of possible futures, and your hypothesis refers to "your view of the underlying forces at work is correct".

(2) How much confidence do you have in the correctness of your understanding of the underlying forces at work? In Bayesian equations, this variable is p (your hypothesis).

(3) Think of the most influential alternative to your perspective on the underlying forces at work. How possible is it that this perspective might be correct? This variable is p (rival hypothesis).

(4) How likely do you think each set of possible futures is if this alternative view of the underlying forces at work is correct? These variables go by the designations p (x1/rival hypothesis), p (x2/rival hypothesis)....

Readers familiar with Bayesian probability theory will immediately recognize these judgments as critical inputs for computing diagnosticity ratios, the likelihood of B given A1 divided by the likelihood of B given A2, and for inferring experts' "priors", or their confidence in competing hypotheses bearing on underlying states of nature. I call these exercises "reputational bets" because they ask experts, in effect, to specify, as exactly as an odds-setting process would have to specify, competing predictions that are predicated on different views of reality. The amount of confidence one should retain in one's prior hypothesis (relative to the most plausible competing hypothesis) after learning what happened is known as the posterior odds. Bayes' theorem tells us precisely how to compute the posterior odds:

$$P\ (\textit{your hypothesis} / X_1\ \textit{occurs}) = P\ (X_1\ /\ \textit{your hypothesis})\ P\ (\textit{your hypothesis})$$

$$P\ (\text{rival hypothesis} / X_1\ \text{occurs})\ P\ (X_1\ /\ \text{rival hypothesis})\ P\ (\text{rival hypothesis})$$

$$\text{Posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

We used this format in forecasts for the Soviet Union (1988), South Africa (1988–89), Canada (1992) and the European Monetary Union (1992).

Our full data set allowed us to answer three categories of questions:

(1) Do experts confronted by new evidence change their minds in the direction and to the approximate degree that Bayes' formula says they should have?
(2) When experts resist changing their minds, what types of justifications (or belief system defenses) do they invoke for holding their ground?
(3) Do systematic individual differences arise in the degree to which experts function like good Bayesian belief updaters?

Figure 12.2 indicates that across all forecasting domains in which we obtained measures of prior probabilities, diagnosticity ratios at the original forecasts, and posterior probabilities at the follow-up session, experts tended to take a "heads-I-win-and-tails-I-do-not-lose" attitude toward forecasting exercises. Experts whose most likely scenarios materialized generally claimed a measure of victory by increasing their confidence in their prior understanding of the underlying forces at work, but a large subgroup of experts whose most likely scenarios failed to materialize denied defeat by showing little inclination to decrease their understanding of the underlying forces at work. Figure 12.2 also shows that experts with

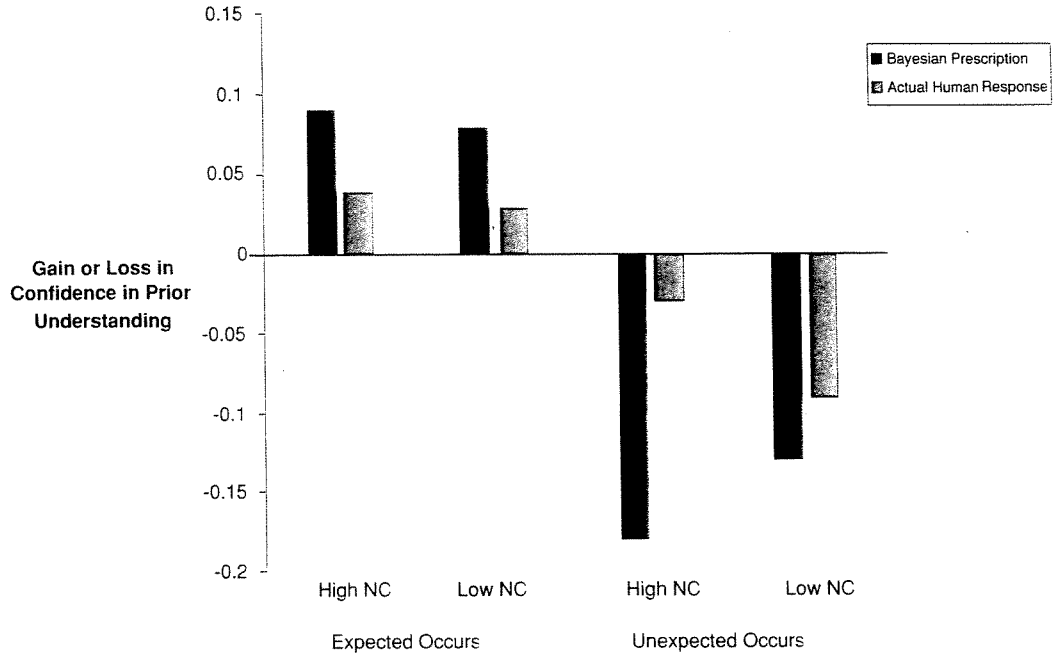

Figure 12.2 This graph shows how much low- and high-need-closure forecasters change their minds in response to expected or unexpected events and compares actual belief updating to the Bayesian-prescribed amount of belief updating

strong preferences for explanatory closure were more prone to resist changing their minds when unexpected outcomes occurred.

On first inspection, these findings would seem to replicate the well-known cognitive-conservatism bias in probabilistic reasoning which asserts that people generally do not change their minds as much as the ideal-type Bayesian would. The term "conservatism" carries, of course, no ideological connotation here, referring not to a particular point of view but rather to conserving existing mental structures or schemata. Liberals can be, and often are, as "guilty" of cognitive conservatism as conservatives.

Caution is, however, in order in drawing conclusions about rationality. Judging the appropriateness of belief updating in the political world is far more problematic than judging it in the classic context of sampling red and blue balls from urns. Experts invoked varying combinations of six strategies for protecting conditional forecasts that ran aground with troublesome evidence. These strategies should not, moreover, be written off as mere psychological defense mechanisms. On close inspection, many turn out to be logically and empirically defensible. We should not fall into the trap of supposing that it is merely a matter of arithmetic to determine whether people are good Bayesians. Some defenses can be viewed, for example, as thoughtful efforts to redefine the terms of reputational bets that, in a Bayesian framework, mandate how much belief change is warranted. The six strategies are as follows.

Strategy 1: The Close-Call Counterfactual Defense (I Was Almost Right)

History provides no control groups. We never know for sure what would have happened if this or that antecedent condition had taken on a slightly different value. Experts often

take advantage of this causal ambiguity to argue that, although the predicted outcome did not occur, it "almost occurred", and it would have indeed occurred but for some inherently unpredictable, seemingly trivial, contingency. Examples of such "close-call counterfactuals" pop up in virtually every forecasting arena in which experts made reputational bets. Consider but the following two cases.

Observers of the former Soviet Union who, in 1988, thought the Communist Party would be firmly ensconced in the saddle of power 5 years hence were especially likely to believe that Kremlin hardliners almost overthrew Gorbachev in the coup attempt of August 1991, as they would have had the conspirators been more resolute and less inebriated, or had key military commanders obeyed orders to kill civilians challenging martial law or had Yeltsin not acted so bravely and decisively.

Experts who expected the European Monetary Union to collapse argued that the event almost happened in the wake of the currency crises of 1992, as, indeed, it would have but for the principled determination (even obstinacy) of politicians committed to the Euro cause and of the interventions of sympathetic central bankers. Given the deep conflict of interest between states that have "solid fundamentals" and those that "regularly resort to accounting gimmickery to make their budget deficits appear smaller", and given the "burbling nationalist resentment" of a single European currency, these experts thought it a "minor miracle" that most European leaders in 1997 were still standing by monetary union, albeit on a loophole-riddled schedule.

Strategy 2: The Just-Off-On-Timing Defense

This strategy moves us out of the murky realm of counterfactual worlds and back into this, the actual, world. Experts often insist that, although the predicted outcome has not yet occurred, it eventually will and we just need to be patient. This defense is limited, of course, in its applicability to political games in which the predicted outcome has not yet been irreversibly foreclosed. No one, for example, expected white-minority rule to be restored in South Africa or Al Gore suddenly to take George W. Bush's place in the White House in 2001. Those were done deals. But it is possible to argue, and experts often did, that a political trend that they deemed highly likely has merely been delayed, and that Canada still will disintegrate (the Parti Québécois will try again and prevail on its third attempt), that Kazakhstan will ultimately burst into a Yugoslav-style conflagration of interethnic warfare (demagogues on both sides of the border with Russia will eventually seize the opportunities for ethnic mobilization that Kazakhstan presents), that the European Monetary Union's misguided effort to create a common currency will some day end in tears and acrimony (the divergent interests of the prospective members will trigger crises that even determined leadership cannot resolve) and that nuclear war will ultimately be the fate of south Asia or the Korean peninsula. In effect, the experts admitted that they may have been wrong within my arbitrary time frames, but they will be vindicated with the passage of time.

Strategy 3: The "I-Made-the-Right-Mistake" Defense

This strategy concedes error, but, rather than trying to minimize the conceptual or empirical significance of the error, it depicts the error as the natural byproduct of pursuing the right

moral and political priorities. As one conservative defiantly declared, "over-estimating the power of the Soviet Union in the 1980s, and the staying power of the Communist Party of the Soviet Union, was the prudent thing to do, certainly a lot more prudent than under-estimating those characters." A mirror-image variant of this defense was invoked by some liberals in the late 1990s in defense of International Monetary Fund (IMF) loans to Russia. Much of the money may have been wasted or misdirected into Swiss bank accounts, but, given the risks of allowing a nuclear superpower to implode financially, issuing the loans was, to paraphrase the second in command at the IMF, Stanley Fischer, the "prudent thing to do".

Strategy 4: Challenge Whether the Preconditions for Activating Conditional Forecasts Were Fulfilled

Each forecast was conditional on the correctness of the expert's understanding of the underlying forces at work. One does not need to be a logician or philosopher of science to appreciate that this is a complex form of "conditionality". Experts have the option of affixing responsibility for forecasting errors on the least ego-threatening and belief-destabilizing mistake that they made in sizing up the situation at the time of the original forecast.

We heard many variants of this refrain in policy debates. One side complains that the other side has unfairly and opportunistically stuck it with an idiotic prediction. The "wronged" side insists, for instance, that they were not mistaken about the efficacy of basic strategies of diplomatic influence, such as deterrence or reassurance, or about the efficacy of basic instruments of macroeconomic policy, such as shock therapy as opposed to gradualism. They merely failed to anticipate how maladroitly the policy would be implemented. Thus, experts could insist at various points between 1992 and 1997 that "if NATO had sent the Serbian leadership the right mix of deterrence-and-reassurance signals, we could have averted further escalation of the Yugoslavian civil war", or "if Yeltsin had practiced real shock therapy, the Russians need not have suffered through this renewed nasty bout of hyperinflation." It is worth noting that this belief-system defense has the net effect of transforming a conditional forecast (if x is satisfied, then y will occur) into an historical counterfactual ("if x had been satisfied, then y would have occurred). Counterfactual history often serves as a conceptual graveyard for conditional forecasts slain by evidence.

Strategy 5: The Exogenous-Shock Defense

All hypothesis testing in science presupposes a *ceteris paribus*, or "all-other-things-equal", clause. Conditional forecasters can thus argue that, although the conditions for activating the forecast were satisfied—their understanding of the underlying forces was correct—key background conditions (implicitly covered by *ceteris paribus*) took on bizarre forms that they could hardly have been expected to anticipate and that short-circuited the otherwise reliably deterministic connection between cause and effect. Theorists tend to be quite comfortable advancing this defense. They can explain away unexpected events by attributing them to plausible causal forces outside the logical scope of their theory. One realist, who was

surprised by how far Gorbachev was prepared to go in making concessions on traditionally sensitive geopolitical and arms-control issues, commented: "I work at the level of relations among states. I am not a specialist on the Soviet Union. You are a psychologist and you should understand the distinction I am making. Would you count it as a failure against a theory of interpersonal relations if the theory predicts that a couple will stay married, the couple stays married for decades, and then suddenly the husband dies of a heart attack. Of course not. The failure, if there is one, lies in failing to check with the cardiologist. Well, my failure, if there was one, was in failing to pay adequate attention to warnings that the Soviet state was very sick."

Strategy 6: The Politics-Is-Hopelessly-Cloud-Like Defense

Finally, experts have the option of arguing that, although the relevant preconditions were satisfied, and the predicted outcome never came close to occurring and now never will, this failure should not be held against the framework that inspired the forecast. Forecasting exercises are best viewed as light-hearted diversions of no consequence because everyone knows, or else should know, that politics is inherently indeterminate, more cloud-like than clock-like. As Henry Kissinger wryly wrote Daniel Moynihan after the fragmentation of the Soviet Union, "Your crystal ball worked better than mine" (Moynihan, 1993, p. 23). On close inspection, of course, this concession concedes nothing.

BIAS NO. 3: HINDSIGHT BIAS

After the specified forecasting periods had elapsed, we asked a subset of experts in six domains to recall their original predictions. The results replicated the core finding in the large research literature on the certainty-of-hindsight effect (Fischhoff, 1975; Hawkins & Hastie, 1990). Experts claimed by an average margin of 0.16 (on a 0–1.0 scale) that they attached higher subjective probabilities to the observed outcomes than they actually did. The results also added the following two wrinkles to the hindsight effect:

(1) The tendency of experts to short-change the intellectual competition. When experts were also asked to recall the predictions they originally thought their most influential rivals would make, they imputed lower conditional likelihoods to the future that materialized than they did prior to learning what happened. In effect, experts claimed to know more about the future than they actually did, and gave less credit to their opponents for anticipating the future than they actually deserved.
(2) The tendency of experts who placed greater value on closure and parsimony to display stronger hindsight effects (cf. Campbell & Tesser, 1983).

BIAS NO. 4: THEORY-DRIVEN STANDARDS OF EVIDENCE AND PROOF

The data contained many examples in which experts applied higher standards of evidence and proof for dissonant than for consonant claims. One striking example was the shifting

pattern of correlates of attitude toward close-call counterfactuals. We saw earlier that experts, especially those who valued explanatory closure, were more favorably disposed toward close-call scenarios that rescued conditional forecasts from falsification (the I-was-almost-right defense). This defense provided a convenient way of arguing that, although the predicted outcome did not occur, it almost did and would have but for theoretically irrelevant twists of fate: "Sure, I thought Canada would have disintegrated by now, and it nearly did", or "I thought South Africa would have lapsed into civil war by now, and it would have but for the remarkable coincidence of two remarkably mature leaders, emerging as leaders at the right moment."

In work on retrospective reasoning, we have found that, although the close-call counterfactual is a welcome friend of the theory-driven conditional forecaster, it is a nuisance at best, and a serious threat at worst to theory-driven thinkers who are on the prowl for ways of assimilating past events into favored explanatory frameworks. Consider the problem of the theorist who subscribes to the notion that the international balance of power is a self-equilibriating system. According to the theory of neorealist balancing, states are unified, rational, decision-making entities that seek to preserve their autonomy; states exist in a fundamentally anarchic interstate environment (no world government) in which, to paraphrase Thucydides, the strong do what they will and the weak accept what they must; therefore, whenever states perceive another state becoming too powerful, they coalesce against it. It was no accident from this point of view that would-be hegemonists—from Philip II of Spain to Napoleon to Hitler—have failed. The military outcomes were the inevitable result of the operation of a basic law of world politics. Not surprisingly, therefore, experts who endorse neorealist balancing theory, and prefer closure and parsimony, are especially likely to reject close-call counterfactuals that imply Napoleon or Hitler could have won if he had made better strategic decisions at various junctures. This theoretical-belief-by-cognitive-style interaction has now been replicated in four distinct conceptual domains (Tetlock & Lebow, 2001).

The most direct evidence for epistemic double-dealing emerges when we present experts with hypothetical discoveries from recently opened archives that either reinforce or undercut favored or disfavored close-call counterfactuals. For example, Sovietologists who subscribed to a monolithic totalitarian image of the Soviet Union cranked up the magnification in looking for flaws in historical work on new Kremlin archives purporting to show that the Communist Party came close to deposing Stalin in the late 1920s and to moving toward a kinder, gentler form of socialism. But the same experts accepted at roughly face value the same historical procedures when the procedures pointed to the more congenial conclusion that Stalinism was unavoidable (even without Stalin). By contrast, Sovietologists who subscribed to a more pluralistic image of the Soviet polity had the opposite reactions (Tetlock, 1999).

BIAS NO. 5: INCOHERENCE AND VIOLATIONS OF EXTENSIONALITY

The term "extensionality" may be forbiddingly technical, but the normative stipulation is simplicity itself. The likelihood of a set of outcomes should equal the sum of the likelihoods of the logically exhaustive and mutually exclusive members of that set. It is hard to imagine a more jarring violation of classic probability theory than the violation of extensionality.

The heuristics-and-biases literature warns us, however, to expect systematic violations of extensionality when people judge complex event sequences that require integrating two or more probabilistic linkages. The textbook illustration is the conjunction fallacy (Tversky & Kahneman, 1983). Imagine that one randomly constituted group is asked to judge the likelihood of a plausible conjunction of events, such as an earthquake causing a dam to rupture that, in turn, causes a flood killing more than 500 people in California. Imagine also that another randomly constituted group is asked to judge the likelihood of a flood (produced by any cause) killing more than 500 people in California. The likelihood judgments of the former group will typically exceed those of the latter group by a substantial margin, even though the former group is judging a subset of the class of outcomes being judged by the latter group.

Building on this work, Tversky and Koehler (1994) and Tversky and Fox's (1995) advanced support theory, which warns us to expect that psychologic will trump logic because people find it easier to mobilize mental support for highly specific possibilities than they do for the abstract sets that subsume these possibilities. For example, people will often judge the likelihood of an entire set of possibilities, such as any NBA team from a given league winning the championship, to be substantially less likely than the sum of the likelihood values that attach to the unpacking of the set's exclusive and exhaustive components (the individual teams that make up the league). The net result is thus that people judge the whole to be less than the sum of its parts (subadditivity) and wind up giving contradictory answers to logically equivalent versions of the same question.

Unpacking manipulations are understandably viewed as sources of cognitive bias on subjective-probability judgments of possible futures. They stimulate people to find too much support for too many possibilities. Returning to the basketball example, Tversky and Fox (1995) demonstrate that although binary complements at the league level generally sum to 1.0 (will the East or West win?), the subjective probabilities assigned to progressively more detailed or unpacked outcomes—the prospects of individual teams within leagues—substantially exceed 1.0. If people were to back up their unpacked bets with actual money, they would be quickly transformed into money pumps. It is, after all, logically impossible that each of four teams within an eight-team division could have a 0.4 chance of winning the championship the same year.

Unpacking manipulations may, however, help to debias subjective probability judgments of possible pasts by exactly the same mechanism. The key difference is that judgments of possible pasts, unlike those of possible futures, have already been contaminated by the powerful certainty-of-hindsight bias. Experimental work on this bias has shown that, as soon as people learn which one of a number of once-deemed possible outcomes happened, they quickly assimilate that outcome knowledge into their existing cognitive structures and have a hard time reconstructing their *ex ante* state of uncertainty (Hawkins & Hastie, 1990). Mental exercises that involve unpacking sets of possible pasts should have the net effect of checking the hindsight bias by bringing back to psychological life counterfactual possibilities that people long ago buried with deterministic "I-knew-it-had-to-be" thinking.

Our research on political experts is consistent with this debiasing hypothesis. In two sets of follow-up contacts with experts on China (1998) and North Korea (1998), randomly-assigned-to-treatment experts were less susceptible to hindsight when they had previously been encouraged to unpack the set of "alternative counterfactual outcomes" and to imagine specific ways in which "things could have worked out very

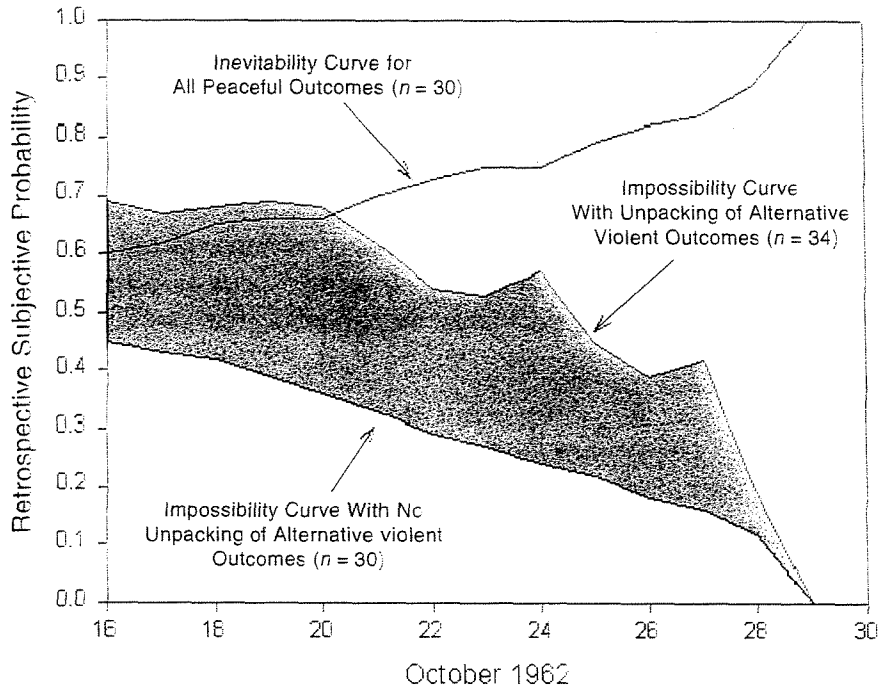

Figure 12.3 Inevitability and impossibility curves for the Cuban Missile Crisis. The inevitability curve displays gradually rising likelihood judgments of some form of peaceful resolution. The lower impossibility curve displays gradually declining likelihood judgments of all possible more violent endings. The higher impossibility curve was derived by adding the experts' likelihood judgments of six specific subsets of more violent possible endings. Adding values of the lower impossibility curve to the corresponding values of the inevitability curve yields sums only slightly above 1.0. Inserting values from the higher impossibility curve yields sums well above 1.0. The shaded area represents the cumulative effect of unpacking on the retrospective subjective probability of counterfactual alternatives to reality

differently" (Tetlock, 2002). Encouraging experts to unpack more temporally distant sets of historical possibilities has also been shown to have a pronounced influence on experts' judgments of the retrospective likelihood of possible outcomes of the Cuban Missile Crisis as well as of other even more remote historical processes. For example, Tetlock and Lebow (2001) report an experiment in which one-half of the participants were asked to imagine the set of alternative more violent endings of the Cuban Missile Crisis and judge the likelihood of the set as a whole on each day of the crisis. The other half were asked to break that set down into exclusive and exhaustive components, including subsets of scenarios in which violence is localized to the Caribbean or extends outside the Caribbean, and further subsets with casualties less than 100 or 100 or more—and then to judge the likelihood of each of these subsets on each day of the crisis. As support theory would lead one to expect, and as Figure 12.3 shows, the experts saw alternative more violent endings of the crisis as significantly more probable when they had performed the decomposition exercise, and this effect was significantly more pronounced among our more "open-minded", low-need-for-closure participants. As can also be inferred from Figure 12.3, when we add the judgments that experts made of the likelihood of some form of peaceful ending on each date (the inevitability curve) to the likelihood of alternative more violent endings taken as a whole set (the wholistic impossibility curve), the sum does not stray too far from 1.0 across dates (the binary-complementarity prediction of support theory). But when we add the points

on the inevitability curve to the corresponding dates on the impossibility curve created by summing subjective probabilities of unpacked what-if scenarios, the sums for the two curves substantially exceed 1.0 on most dates (consistent with the subadditivity prediction of support theory).

Here, we confront another normative judgment that looks easy in laboratory settings but more problematic in the real world. From a strictly logical point of view, subadditivity is indefensible. If we believe, however, that historical reasoning is already biased by distortions of hindsight, there is a good case that encouraging experts to imagine counterfactual alternatives to reality, and inflating their subjective probability estimates beyond the bounds of reason, might be a reasonable thing to do if it checks the hindsight bias.

SOME CLOSING OBSERVATIONS

How should we balance these potentially endless arguments bearing on the rationality of professional observers of world politics? It is useful here to draw a sharp distinction between the descriptive and the normative; between the generalizability of purely empirical characterizations of particular judgmental tendencies and the generalizability of the normative characterizations of those judgmental tendencies as errors or biases. The studies reviewed in this chapter attest simultaneously to the empirical robustness and the normative contestability of error-and-bias claims in the hurly-burly of world politics. The case for empirical robustness is strong. Consider the following seven examples of how the real-world evidence reported here converges with the laboratory evidence in the mainstream literature:

(1) The overconfidence documented in political forecasts reaffirms a massive body of work on the calibration of subjective probability estimates of knowledge (Dawes, 1998).
(2) The selective activation of belief-system defenses by forecasters who "get it wrong" dovetails nicely with the classic dissonance prediction that people would most need defenses when they appear to have been wrong about something on which they were originally quite confident (Festinger, 1964).
(3) The skepticism that experts reserved for dissonant historical evidence and claims extended the work on theory-driven assessments of evidence and on the tendency for people to apply stringent "must-I-believe" tests to disagreeable evidence and much more lenient "can-I-believe" tests to agreeable discoveries (Griffin & Ross, 1991).
(4) Experts' generation of close-call counterfactuals in response to unexpected events is consistent with experimental work on norm theory and the determinants of spontaneous counterfactual thinking (Kahneman & Miller, 1986).
(5) The reluctance of experts to change their minds in response to unexpected events and in accord with earlier specified diagnosticity ratios parallels the excessive conservatism in belief revision often displayed by subjects in experiments that explicitly compare human judgment to Bayesian formulas (Edwards, 1968).
(6) The cognitive-stylistic differences in belief-system defense and belief underadjustment offer further evidence for the construct validity of the need-for-closure and integrative complexity measures (Kruglanski & Webster, 1996; Suedfeld & Tetlock, 2001).
(7) The subadditivity effects induced by encouraging experts to unpack counterfactual alternatives to reality is consistent both with Tversky's support theory and with work on the power of imagining alternative outcomes to check hindsight bias (Koehler, 1991).

In all seven respects, the current results underscore the generalizability of laboratory-based demonstrations of bounded rationality in more ecologically representative research designs. The psychological findings do hold up well when highly trained experts (as opposed to sophomore conscripts) judge complex, naturally occurring political events (as opposed to artificial problems that the experimenter has often concocted with the intent of demonstrating bias).

Curiously, though, empirical robustness coexists with normative contestability in the political realms surveyed here. Why should normative judgments of rationality become so much more problematic when the object of inquiry is political in content and historical in process? The answer appears to be at least threefold:

1. Politics is often defined as an organized competition for power in which rival communities of cobelievers warn of looming threats and advocate particular policies to avert threats and bring about consequences that most people deem desirable (for example, "We predict race riots if we don't adopt more egalitarian policies"; "We predict aggression by expansionist/revisionist states if we don't adopt stronger deterrence policies"). Therefore, it should not be surprising when experts representing competing theoretical or ideological camps place different values on avoiding type I as opposed to type II errors in predicting various outcomes. What looks like overconfidence within one camp will frequently look like a prudent effort to minimize the really serious type of error from the standpoint of the other camp.

2. Inasmuch as political cognition tends to occur in a highly adversarial environment, we should expect the players to be keenly aware that the other side will be prepared to pounce on potentially embarrassing errors. This helps to explain why it is difficult to persuade experts to make falsifiable forecasts even when they have been explicitly assured that all judgments will be absolutely confidential. Many participants in our studies work within professional cultures in which their reputations hinge on appearing approximately right most of the time and on never appearing clearly wrong. What looks like an error or bias from a Bayesian standpoint (an unwillingness to stick to reputational bets made earlier) can also be plausibly viewed as a strategic adaptation to the rhetorical demands of thrust and parry in highly partisan contest for power. As one expert told me, "You think we're playing the (hypothetico-deductive) game of science, and so you evaluate what is going on by those standards. But that is as silly as trying to apply the rules of football to baseball. In the game of politics, truth is secondary to persuasion."

3. Even granting this objection, the naive behavioral scientist might still wonder whether it is possible for rival communities of cobelievers to insulate themselves from falsification indefinitely. Surely, "truths"—to which everyone must be responsive—will slowly become undeniably apparent. This counterargument does not, however, adequately take into account the profound obstacles that arise in assessing historical causation. Sharp disagreements still exist over why World War I or, for that matter, the English Civil War of the 1640s broke out when and in the manner it did, and whether it could have been averted by this or that counterfactual alteration. It is not unusual for these sorts of disputes to persist for centuries, and even millennia (Tetlock, 2002).

Disagreements over causation are so intractable largely because all causal inference in history ultimately rests on speculative counterfactual judgments of how events would have unfolded if this or that antecedent condition, hypothesized to be relevant by this or that camp, had taken on a different value (Fogel, 1964; Fearon, 1991). The political observers

in the current studies confronted the daunting task of making sense of constantly evolving path-dependent sequences of events with indeterminate numbers of branching points. The traditional scientific methods of causal inference—experimental and statistical control—just do not apply. Experimental control was not an option because history is a path-dependent system that unfolds once and only once. Statistical control was not an option because of the well-known problems of classifying complex, highly idiosyncratic events that many experts insist are categorically unique (and hence resistant to all classification), and that the remaining experts often insist on assigning to incompatible classificatory bins. Testing hypotheses about the effects of fuzzy-set concepts such as "deterrence" or "democracy" on war proneness requires, at minimum, agreement on when deterrence was or was not implemented, and on whether a given state qualifies as democratic.

In brief, our experts typically worked under loose reality constraints that made it easy to wriggle out of disconfirmation. We saw, for example, even with *ex ante* likelihood ratios in hand, how extraordinarily difficult it was to make a decisive case that any given individual was guilty of biased information processing or belief perseverance. History permits of too many explanations. Going backward in time, political partisans could always argue that they were not almost wrong, and, going forward in time, they could always insist that they were almost right. And who is to say for sure that anything is amiss. No one can visit these counterfactual worlds to determine which what-if assertions are defensive nonsense, and which ones are on target.

Although making accusations of irrational belief perseverance logically stick is extremely difficult, the studies reported here do still reveal ample grounds for concern that many political debates are equivalent to Einhorn and Hogarth's (1981) "outcome-irrelevant learning situations". Loose reality constraints coupled with the human propensity to theory-driven thinking make it easy for even sophisticated political observers to slip into tautological patterns of reasoning about history that make it well-nigh impossible for them ever to discover that they were wrong. For this reason, this chapter advances the argument that it is a mistake to treat subadditivity in judgments of alternative worlds as just a logical error; it is an error that can be put to good use in checking excessively theory-driven modes of making sense of history. We have seen that these theory-driven patterns of reasoning about historical outcomes, especially temporally distant ones, tend to be convergent. The focus is typically on explaining why what was had to be. The subadditivity effects appear, however, to be the product of divergent, imagination-driven thinking. The focus is (at least in the studies reported here) on what could have been. The theory-driven strategies confer the benefits of explanatory closure and parsimony by assuring us that we now know why things worked out as they did. But these strategies desensitize us to nuance, complexity and contingency. The imagination-driven strategies sensitize us to possible worlds that might or could have been, but the price can be increased confusion, self-contradiction and even incoherence.

One of the deepest conceptual challenges in historical reasoning may be that of striking a reasonable balance, a reflective equilibrium, between convergent theory-driven thinking and divergent imagination-driven thinking. On the one hand, historical observers need imagination-driven modes of thinking to check the powerful tendency to assimilate known outcomes into favorite causal schemata. On the other hand, observers need theory-driven modes of thinking to check runaway unpacking effects and to serve as plausibility pruners for cutting off speculation that would otherwise grow, like Topsy, beyond the bounds of

probability. Of course, there is no single, well-defined equilibrium or normative solution. So there will be plenty of room for competing communities of cobelievers to stake out different theoretical and ideological standards for what counts as a reasonable balance. It is therefore a safe bet that setting standards of good political judgment will continue to be politically controversial.

REFERENCES

Campbell, J. D. & Tesser, A. (1983). Motivational interpretations of hindsight bias: an individual difference analysis. *Journal of Personality, 51*, 605–620.

Dawes, R. (1997). Judgment and choice. In D. Gilbert, S. Fiske & G. Lindzey (eds), *Handbook of Social Psychology*. New York: McGraw Hill.

Dawes, R. (1998). Judgment and Choice. In D. Gilbert, S. Fiske & G. Lindzey (eds), *Handbook of Social Psychology (Volume 1)* (pp. 497–548). New York: McGraw Hill.

Einhorn, H. & Hogarth, R. (1981). Behavioral decision theory: processes of judgment and choice. *Annual Review of Psychology, 31*, 53–88.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (ed.), *Formal Representation of Human Judgment* (pp. 17–52). New York: Wiley.

Elster, J. (1978). *Logic and Society: Contradictions and Possible Worlds*. New York: Wiley.

Fearon, J. (1991). Counterfactuals and hypothesis testing in political science. *World Politics, 43*, 474–484.

Festinger, L. (ed.) (1964). *Conflict, Decision, and Dissonance*. Stanford, CA: Stanford University Press.

Fischhoff, B. (1975). Hindsight is not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology, 104*, 288–299.

Fogel, R. (1964). *Railroads and American Economic Growth: Essays in Econometric History*. Baltimore, MD: Johns Hopkins University Press.

Friedman, T. (1999). *The Lexus and the Olive Tree*. New York: Farrar, Straus, & Giroux.

Goldstein, W. & Hogarth, R. (eds) (1996). *Judgment and Decision Making: An Interdisciplinary Reader*. Cambridge: Cambridge University Press.

Griffin, D. & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In M. Zanna (ed.), *Advances in Experimental Social Psychology* (volume 24, pp. 319–359). New York: Academic Press.

Hawkins, S. & Hastie, R. (1990). Hindsight: Biased judgment of past events after outcomes are known. *Psychological Bulletin, 107*, 311–327.

Kahneman, D., Slovic, P. & Tversky, A. (eds) (1982). *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Kahneman, D. & Miller, D. (1986). Norm theory: comparing reality to its alternatives. *Psychological Review, 93*, 136–153.

Koehler, D. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin, 110*(3), 499–519.

Kruglanski, A. & Webster, D. (1996). Motivated closing of the mind: seizing and freezing. *Psychological Review, 103*, 263–278.

Moynihan, D. P. (1993). *Pandemonium*. New York: Oxford University Press.

Nisbett, R. E. & Ross, L. (1980). *Human Inference:Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice–Hall.

Slovic, P., Fischhoff, B. & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28*, 1–39.

Suedfeld, P. & Tetlock, P. E. (1999). Cognitive styles. In A. Tesser & N. Schwartz, *Blackwell International Handbook of Social Psychology: Intra-individual Processes* (vol 1 •••). London: Blackwell.

Suedfeld, P. & Tetlock, P. E. (2001). Cognitive styles. In A. Tesser & N. Schwartz, *Blackwell International Handbook of Social Psychology: Intra–individual Processes, (Vol 1)* (pp. 284–304). London: Blackwell Publishers.

Tetlock, P. E. (1991). Learning in U.S. and Soviet foreign policy: in search of an elusive concept. In George Breslauer & Philip Tetlock (eds), *Learning in U.S. and Soviet Foreign Policy*. Boulder, CO: Westview.

Tetlock, P. E. (1992). Good judgment in world politics: three psychological perspectives. *Political Psychology, 13*, 517–540.

Tetlock, P. E. (1998). Close-call counterfactuals and belief system defenses: I was not almost wrong but I was almost right. *Journal of Personality and Social Psychology, 75*, 230–242.

Tetlock, P. E. (1999). Theory-driven reasoning about possible pasts and probable futures: are we prisoners of our preconceptions? *American Journal of Political Science, 43*, 335–366.

Tetlock, P. E. (2002). Social-functionalist metaphors for judgment and choice: the intuitive politician, theologian, and prosecutor. *Psychological Review, 109*, 451–472.

Tetlock, P. E. & Belkin, A. (1996). *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton, NJ: Princeton University Press.

Tetlock, P. E. & Lebow, R. N. (2001). Poking counterfactual holes in deterministic covering laws: Alternative histories of the Cuban missile crisis. *American Political Science Review, 95*(4), 829–843.

Tversky, A. & Fox, C. (1995). Weighting risk and uncertainty. *Psychological Review, 102*(2), 269–83.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reason: the conjunction fallacy as probability judgment. *Psychology Review, 90*(2), 292–315.

Tversky, A. & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.

Winkler, R. L. (1994). Evaluating probabilities: asymmetric scoring rules. *Management Scence,40*, 1375–1405.