Theory-Driven Reasoning About Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions?

Philip E. Tetlock

American Journal of Political Science, Vol. 43, No. 2 (Apr., 1999), 335-366.

Stable URL:
http://links.jstor.org/sici?sici=0092-5853%28199904%2943%3A2%3C335%3ATRAPPA%3E2.0.CO%3B2-K

Theory-Driven Reasoning about Plausible Pasts and Probable Futures in World Politics: Are We Prisoners of Our Preconceptions?

Philip E. Tetlock, *The Ohio State University*

Cognitive theories predict that even experts cope with the complexities and ambiguities of world politics by resorting to theory-driven heuristics that allow them: (a) to make confident counterfactual inferences about what would have happened had history gone down a different path (plausible pasts); (b) to generate predictions about what might yet happen (probable futures); (c) to defend both counterfactual beliefs and conditional forecasts from potentially disconfirming data. An interrelated series of studies test these predictions by assessing correlations between ideological world view and beliefs about counterfactual histories (Studies 1 and 2), experimentally manipulating the results of hypothetical archival discoveries bearing on those counterfactual beliefs (Studies 3–5), and by exploring experts' reactions to the confirmation or disconfirmation of conditional forecasts (Studies 6–12). The results revealed that experts neutralize dissonant data and preserve confidence in their prior assessments by resorting to a complex battery of belief-system defenses that, epistemologically defensible or not, make learning from history a slow process and defections from theoretical camps a rarity.

The now voluminous literature on judgmental biases and errors raises doubts about the capacity of even highly motivated professionals to perform the types of information-processing tasks that rational-actor models regularly posit people routinely perform (cf. Dawes, 1998; Kahneman, Slovic, and Tversky, 1982; Hogarth and Goldstein, 1996). This largely cognitive-psychological literature casts doubt on our ability to draw sound causal conclusions from complex arrays of data, to reason consistently about the trade-offs implied by multiattribute choice tasks, to attach appropriate confidence to predictions, and to revise beliefs in a timely fashion in response to new evidence.

The author acknowledges the financial support of the Institute of Personality and Social Research at the University of California, Berkeley, the Center for Advanced Study in the Behavioral Sciences, the Mershon Center of The Ohio State University and the MacArthur Foundation. The author also thanks the following persons for helpful advice and assistance: Daniel Kahneman, Richard Herrmann, Terry Busch, Mary O'Sullivan, Richard Boettger, Aaron Belkin, Beth Elson, Charles McGuire, Greg Mitchell, Sara Hohenbrink, and Rachel Szteiter. Finally, the author owes the greatest debt to the many thoughtful professionals who made time in their busy schedules to grapple with a battery of exceedingly difficult questions. I harbor no illusions that I could have done a better job than they.

American Journal of Political Science, Vol. 43, No. 2, April 1999, Pp. 335–366

These "cognitivist" challenges to rationality are themselves open to methodological and conceptual challenges. The most frequently recurring counter-challenge has been to impugn the external validity of the largely laboratory-based sources of evidence on judgmental bias. Skeptics often argue that laboratory researchers rely far too heavily on inexperienced undergraduate subjects, confront subjects with artificial tasks, fail to motivate subjects with adequate incentives, and ignore the social and institutional constraints present in most everyday situations (cf. Kagel and Roth, 1995; Tetlock, 1992a).

Researchers have responded to this barrage of criticisms by demonstrating that many "deviations from rationality" hold up even when highly trained professionals confront inferential tasks that are, in key respects, representative of those that professionals confront in their working lives. Meteorologists, radiologists, subatomic physicists, stock market analysts, financial forecasters, and air traffic controllers have all become targets of cognitive investigation (Dawes, 1998). Researchers do, however, pay a price when they leave the laboratory to study "real people" working on "real problems." The typical laboratory paradigm offers a clear-cut normative benchmark for defining response tendencies as errors or biases. For example, if we ask participants to assume that there are fifty red and fifty blue marbles in an urn and then inform them that ten marbles were randomly sampled from the urn and that eight were red and two were blue, Bayes theorem tells us precisely the extent to which participants should adjust their prior beliefs abut the proportion of differently colored marbles. We lose this precision when we gauge the reactions of real-world professionals responding to real-world events—say, the reactions of market analysts to shifting macroeconomic indicators. There is uncertainty about the right answers, the "diagnosticity" of the data is open to debate, and third-party attributions of irrational belief perseverance and overconfidence to analysts become contentious.

The studies reported here explore conceptual and empirical issues that arise in assessing the applicability of the error-and-bias literature to professional observers of world politics. The first set of studies examine the tightness of the connection between experts' conceptions of what would have happened in "counterfactual worlds" and their general ideological outlook. The key question is the extent to which counterfactual reasoning about major historical events is theory-driven (predictable from abstract preconceptions) as opposed to data-driven (constrained by the peculiar "fact situation" of each historical episode). The second set of studies examine the willingness of experts to revise basic counterfactual beliefs in response to archival discoveries that shed new light on how events might have unfolded if assorted historical background conditions had been altered. The key question here is the extent to which experts apply the same standards of evidence and

proof to ideologically dissonant and consonant discoveries. The third set of studies shift the focus from judging possible pasts to probable futures. These studies are unique in that they longitudinally track the reactions of experts to the apparent confirmation or disconfirmation of conditional forecasts of real-world events in real time. The key questions revolve around: (a) the extent to which experts who "get it wrong" resort to various belief-system defenses; (b) the effectiveness of these defenses in helping experts who get it wrong to preserve as much confidence as those who get it right in the underlying correctness of their world views.

Methodologically, the current studies follow in the tradition of naturalistic cognitive psychology. The empirical spotlight is on "real people" trying to solve problems that, if not exactly representative of what they confront on a daily basis, bear a deep structural resemblance to those tasks. Professional observers of world politics are frequently asked to extract lessons from history that require making counterfactual assumptions about alternative paths history might have taken but for this or that contingency. They must frequently judge the probativeness of evidence bearing on counterfactual claims and counterclaims. And they are frequently asked what the future might hold in general or conditional upon various decisions being made. To be sure, the primal epistemological trade-off still applies: what we gain in "ecological representativeness" we lose in ability to draw precise conclusions about when experts have gone too far in allowing their preconceptions to color their interpretations of new observations. There is no sharp dividing line between rationality and irrationality, the defensible and the indefensible, in the real world. But, that said, the results reported here are not devoid of normative implications. They speak directly to how easy or difficult it is for experts to concede that they were wrong about either possible pasts or probable futures, about what might have been or what might yet be.

1. Theory-Driven Inferences about Counterfactual Worlds: Studies I and II

Learning from history is hard, in part, because history is a terrible teacher. It provides no control groups that students can use to test directly what would have happened if Archduke Ferdinand had not been assassinated in 1914 (might World War I have been averted?), if Hitler had perished in a traffic accident in 1930 (might World War II have been averted?), or if Kennedy had heeded his hawkish advisors during the Cuban missile crisis (might World War III have been triggered?) The control groups "exist"—if "exist" is the right word—only in the more or less disciplined imaginations of the observers (Fearon, 1991; Tetlock and Belkin, 1996).

But the students may not be blameless. Learning from history may also be hard because even professionals are cognitively ill-equipped to cope with

the complexity, ambiguity, and dissonance inherent in the task. Research on belief perseverance suggests that observers are often too quick to fill in the missing control conditions of history with elaborate narrative sequences (scripts) that reflect deep-rooted ideological assumptions about both political causality and the characters of specific political actors (Griffin and Ross 1991; Jervis, 1976). From this standpoint, it is not surprising that many conservative advocates of deterrence were convinced that the Cold War would have lasted longer than it did in a hypothetical world with a two-term Carter presidency (e.g., Pipes 1993) whereas many liberal advocates of reassurance were equally persuaded that the Cold War would have ended largely as it did under the Reagan presidency (e.g., Garthoff 1994). Moreover, there need be nothing intellectually dishonest in this striking coincidence between preconceptions and outcomes of counterfactual thought experiments. The results of the thought experiments flow naturally from the causal schemata that observers rely upon for organizing an otherwise unmanageably complex array of historical events. Whatever their logical status, phenomenologically speaking, counterfactual beliefs often feel factual. People express levels of confidence in controversial counterfactuals that are frequently indistinguishable from those they express in events that they have directly witnessed.

Studies I and II explore the balance that experts strike in their counterfactual reasoning between theory-driven information processing (in which the conclusions experts draw from what-if thought experiments are completely predictable from their ideological world view) and data-driven information processing (in which experts are highly sensitive to the idiosyncratic nuances of each historical case and there is little predictability in their responses across cases). Two-stage models of counterfactual reasoning proposed by Roese and Olson (1995) and Tetlock (1998) predict that: (a) more theory-driven forms of thinking will dominate judgments of antecedent-consequent linkages (assuming x took this form, then y would have occurred); (b) more data-driven forms of thinking will dominate judgments of the plausibility of supposing that antecedent historical conditions could have taken on forms different from those they took on in this world.

1.1 Study I: Judging Possible Pasts of the Soviet Union

This study was conducted in 1992 and asked fifty-two specialists on the former Soviet Union (a sample that consisted of M.A. and Ph.D. level professionals working in government and academia) to judge the plausibility of seven counterfactuals that explored contested choice points in Soviet history from the Bolshevik Revolution to the disintegration of the state in 1991—the types of counterfactuals that, as Breslauer (1996) has insightfully documented, have long divided the Sovietological community along ideological

and philosophical lines.[1] For each assertion, experts judged on nine-point scales: (1) the difficulty of imagining the antecedent condition occurring instead of what actually happened; (2) assuming for sake of argument that the antecedents did occur, the likelihood of the hypothesized consequence. Experts also responded to a five-item ideology scale (Cronbach's alpha for elite samples = .89) presented in the appendix. Scores on the scale were normally distributed around a mean of 5.96 (indicating a moderately liberal peak in the ideology distribution).

As Table 1 reveals, ideological disagreements over antecedent-consequent linkages in counterfactual arguments are the rule rather than the exception. Conservatives are more likely than liberals to accept that the Bolshevik revolution could have been avoided but for the chaos of Russia's defeat in World War I and that Soviet foreign policy in the late 1980's would not have been nearly so accommodative if Reagan had not taken so tough a stand in the early 1980's. Conversely, liberals are more likely than conservatives to believe that a kinder gentler communism could have emerged earlier not only by purging Stalin but also by adding 10 years to Lenin's life or by allowing Malenkov to prevail in the post-Stalin succession struggle. Liberals are also more likely to suspect that Gorbachev was critical in steering the Soviet Union down a reformist path and that Gorbachev could have held together some form of Soviet federation if he had been a shrewder tactician.

Inspection of correlation coefficients and of thought protocols lends support to a two-stage cognitive model of processing counterfactual arguments (cf. Roese and Olsen, 1995; Tetlock, 1998). Ideology is a much more consistent predictor of judgments of conditional linkages (if X, then Y) than of the mutability of antecedent conditions (is it reasonable to imagine X taking on a different value?), and this differential predictability is not due to restriction-of-range artifacts (tests for heterogeneity of variance such as Hartley's F max test yielded only two borderline significant differences across the two classes of dependent variables). There was, however, one conspicuous exception to this generalization. Liberals and conservatives strongly disagreed on the plausibility of both the antecedent and the conditional linkage for the "Stalinism" counterfactual. Conservatives had a harder time than did liberals imagining that the Soviet Communist Party could have purged or would have wanted to purge Stalin at this juncture. From an essentialist totalitarian perspective on the Soviet Union, the deletion-of-Stalin counterfactual violates the "minimal-rewrite-rule" (Tetlock and Belkin 1996), which stipulates that sound counterfactuals should do as little violence to the historical record as

[1]An appendix, available on the *AJPS* website (http://psweb.sbs.ohio-state.edu/ajps/), lists these counterfactual claims as well as the response scales used.

Table 1. Correlations Between Political Ideology and Counterfactual Beliefs of Area-Study Specialists

Counterfactuals About Soviet Union	Antecedent	Antecedent/Consequent Linkage
No WWI, no Bolshevik revolution	.25	–.57
Longer life to Lenin, no Stalinism	.13	.68
Depose Stalin, kinder, gentler communism	.66	.70
Malenkov prevails, early end to Cold War	.17	.71
No Gorbachev, CPSU moves in more conservative direction	–.16	.30
No Reagan, no early end to Cold War	–.30	–.74
A shrewder Gorbachev, Soviet Union survives	.11	.51
About South Africa		
No de Klerk, still white-minority rule	.15	–.42
No Mandela, still white- minority rule	.08	.10
No Western sanctions, still white-minority rule	.06	.48
No demographic pressures, still white-minority rule	.11	.15
No Soviet collapse, fewer white concessions	.18	–.51

possible. But from a more pluralistic perspective on the Soviet polity (Cohen, Rabinowitch, and Sharlet 1985), the counterfactual may well pass this test. Second, liberals and conservatives disagree on what would have happened if Stalin had been deposed. Like most counterfactuals, this one is elliptical. It does not spell out the complex connecting principles necessary for bridging the logical gap between antecedent and consequent. To hold the counterfactual together, it is necessary to posit that latent advocates of Gorbachev-style socialism in the CPSU would have seized the reins of power and guided

the Soviet state toward social democracy. Conservatives—who view the Bolshevik Party of the time as monolithically oppressive—regard such arguments as hopelessly contrived.

In sum, ideology predicts the perceived feasibility of major reroutings of history as long as experts assume the mutability of the antecedent; but ideology emerges as a major predictor of the mutability of only one antecedent, Stalin, that tapped not only into philosophical assumptions about historical contingency but also into politically charged assumptions about Bolshevism.[2]

1.2 Study II: Judging Possible Pasts for South Africa

The counterfactual judgments of Soviet history are open to two interpretations. One hypothesis is that those on the left view history in general as more fluid, contingent, and indeterminate than do those on the right. A harsher variant of the same hypothesis depicts more conservative observers as more prone to the certainty-of-hindsight bias (Fischhoff 1975)—the tendency to slip into viewing what happened as retrospectively inevitable by quickly forgetting how uncertain they once were about what would happen. An alternative hypothesis is that liberal and conservative experts reason in fundamentally similar ways, but there is something special about the Soviet Union that motivates wistful perceptions of "lost possibilities" on the left and angry accusations of "inevitable repression and expansion" on the right. If we could identify a state that excites comparable fear and loathing on the left to that once excited by the Soviet Union on the right, we would observe a sign reversal of the correlation coefficients between ideological sympathies and counterfactual beliefs. Those on the left would now reject the possibility that accidents of health or assassination could have redirected a profoundly oppressive system onto the path of reform but embrace the possibility that, absent relentless external pressure, reform never would have occurred.

South Africa seems the ideal case for teasing these hypotheses apart. Accordingly, in 1995 we asked a sample of twenty-five academics, policy analysts in think-tanks, journalists, and intelligence analysts in government (all of whom had written on South Africa) to judge the five counterfactual propositions listed in the appendix available.

Table 1 shows that political ideology once again predicted numerous antecedent-consequent linkages. Conservatives assigned more credit to

[2]An ultimately philosophical question is whether we should think of ideology as exerting a causal influence on counterfactual beliefs or think of counterfactual beliefs as integral logical components of an ideological world view. There is, of course, no hard and fast rule for deciding how empirically correlated and conceptually overlapping two constructs must be before the relationship is classified, in Kantian terms, as analytic rather than synthetic.

de Klerk and to the collapse of Soviet-style communism whereas liberals assigned more credit to Western economic sanctions. There was, however, little ideologically-grounded disagreement over the impact of Mandela and of demographic pressures which both ideological camps deemed important. There was also little disagreement about what Kahneman and Miller (1986) call the mental mutability of antecedents, with consensus that it was easy to subtract particular individuals from the historical matrix, difficult to undo Western sanctions, and very difficult to reverse the demographic pressures.

The openness of conservatives to the de Klerk counterfactual in the South African case roughly parallels the openness of liberals to the Stalin, Malenkov, and Gorbachev counterfactuals in the Soviet case; liberal skepticism toward the Reagan-pressure counterfactual in the Soviet case roughly parallels conservative skepticism toward the economic-sanctions counterfactual in the South African case. But the parallels are imperfect. Observers of South Africa do at least agree on one counterfactual that pivots on a particular person: Mandela.

Taken as a whole, the South African data undermine the notion that liberals subscribe to an inherently more indeterminate and contingency-laden philosophy of history than do conservatives. Openness to counterfactuals is less a function of general ideological orientation and more a matter of the specific sympathies and antipathies activated by particular problems.

The data from the first two studies are strikingly consistent with the oft-quoted observation that people use the past to prop up their prejudices. The data also underscore how difficult it is to prevent the wholesale politicization of knowledge claims in world politics when all causal inference hinges on counterfactual assumptions that, in turn, are guided by ideological preconceptions. What is to stop us from positing counterfactuals of convenience that justify whatever causal assertions we find it expedient to make? Tetlock and Belkin (1996) attempt to answer this question by specifying criteria that scholars can deploy to winnow out specious counterfactuals. Here, however, it must suffice to note that the epistemological problems are daunting but not overwhelming. Although counterfactual claims are strictly speaking about empirically inaccessible "possible worlds," it is often possible to derive testable implications from such propositions that can be investigated in this world. For example, historians can shed light on how close the CPSU ever came to dismissing Stalin by documenting latent resistance to his leadership as well as by exploring the views of his potential successors. But there is still a cognitive catch: to prevent counterfactual speculation from sliding into the solipsistic abyss, experts must be willing to change their minds about possible worlds in response to real-world evidence. As we shall now see, many seem reluctant.

2. Theory-Driven Standards of Evidence in Judging Counterfactual Scenarios: Studies III–V

Counterfactual claims about history resist direct test. No one can hop into a time machine, travel back to undo a key event, and then document what happened. But this impossibility does not imply that evidence is irrelevant. There is a voluminous literature on the logical, theoretical, statistical, and historical criteria that scholars should use in judging counterfactuals (Elster 1978; Tetlock and Belkin 1996). Most of us suspect that some counterfactuals are more compelling than others, and this literature suggests that the intuition may often be defensible.

Consider the sharply contested counterfactual claim that "if the CPSU had deposed Stalin in the late 1920's, the U.S.S.R. would have moved toward a kinder, gentler form of communism 50 years earlier." Sovietologists might reject this claim either on the ground that the antecedent is preposterous or on the ground that even if we concede the feasibility of deposing Stalin, the alternative would not have been better because terror was integral to the logic of the Leninist regime. Let us posit, however, a thought experiment in which historical sleuths in the Kremlin archives claim to discover documents that reveal rising resistance to Stalin in the late 1920's and that, given the chance, the most likely successors would have constructed a kinder, gentler communism. How should experts respond? It seems a trifle dogmatic to refuse even to consider changing one's mind. But such a response might be justifiable if overwhelming evidence from credible sources pointed to the contrary conclusion. Many scientists justify their dismissal of experimental evidence of extrasensory perception on the ground that such findings violate too many well established physical and biological laws. In Bayesian terms, it would be presumptuous for nonexperts to tell experts how "diagnostic" particular evidence is with respect to particular causal hypotheses.

It is possible, however, to design a better mousetrap for documenting the impact of theory-driven thinking about counterfactual history. Imagine that we transform our thought experiment into an actual experiment that holds the evidence constant—say, documents recently discovered in the Kremlin archives—but manipulates the direction of the findings—say, whether the documents contain revelations favorable either to those who view Stalinism as a tragic aberration or to those who view it as a logical outgrowth of Leninism. Insofar as observers deem evidence probative *only* when it reinforces their prior beliefs, the experiment would reveal a disturbing double standard in which judgments of diagnosticity are driven solely by the consistency of the evidence with preconceptions, not by the rigor of the research procedures. To the degree this is so, there is a risk that the observer's beliefs about historical causality are empty tautologies in which preconceptions determine

beliefs about counterfactual worlds that, in turn, shape evaluations of evidence bearing on those beliefs.

Experimental work on social cognition suggests that people often have lower standards of evidence for congenial claims (Fiske and Taylor 1991; Griffin and Ross 1991). This section reports three interrelated studies that replicate and extend this basic effect in the context of expert reasoning about evidence bearing on contested historical counterfactuals. The research paradigm involved transforming thought experiments into actual experiments by asking respondents how they *would* react *if* a research team working in the Kremlin archives announced the discovery of evidence that shed light on three choice points in Soviet history: whether Stalinism could have been averted in the late 1920's, whether the Cold War could have been brought to an earlier end in the mid 1950's, and whether the Politburo in the early 1980's could just as easily have responded to Reagan's policies in a confrontational as opposed to an accommodationist manner. Respondents included the same sample of forty-seven professionals used in Study I plus a sample of twenty-eight experts from both think tanks and universities.

The appendix presents the details of the research design which took the form of a $2 \times 2 \times 3$ mixed-design factorial, with two between-subjects independent variables—liberal or conservative tilt of evidence discovered by hypothetical research team and the presence or absence of methodological checks on ideological bias—and one counterbalanced repeated-measures replication factor based on the three historical "discoveries." Participants were randomly assigned to the 2×2 components of the design. In the liberal-tilt condition, participants imagined that a research team uncovers evidence in Kremlin archives that indicates history could easily have gone down different paths at three junctures: specifically, Stalinism was avertable in the late 1920's, it was possible to end the Cold War in mid-1950's, and Reagan almost triggered a conflict spiral in American-Soviet relations in the early 1980's. In the conservative-tilt condition, participants imagined the discovery of the same types of evidence, but the evidence now indicates that history could not have gone down a different path at each of these three junctures. In the high-research-quality condition, participants are further asked to imagine that the research team took special precautions to prevent political bias and to consider alternative explanations. In the unspecified-research-quality condition, participants received no such assurances. After reading about each discovery, participants rated the credibility of the research team's conclusions and the degree to which they endorsed three distinct grounds for impugning the team's credibility (dismissing the motives of the research team as political rather than scholarly, disputing the authenticity of documents, and arguing that key documents have been taken out of context).

Table 2 shows that, across scenarios, implementing methodological precautions had little impact on the "believability" of the research report. Regardless of announced checks on bias, both liberals (M's = 7.1 versus 7.0) and conservatives (M's = 7.0 versus 6.8) rated consonant evidence as highly credible and dissonant evidence as relatively incredible (for liberals, M's = 4.3 versus 3.7; for conservatives, M's = 4.5 versus 3.2). Repeated-measures analysis of variance revealed the interactive effect of political ideology and evidence-tilt to be highly significant, with liberals rating the "liberal results" as much more credible than the "conservative results", M's = 7.05 versus 4.0, F (1, 67) = 12.97, $p < .001$, and conservatives doing exactly the opposite, M's = 6.9 versus 3.9, F (1, 67) = 12.61, $p < .001$. No other interaction attained significance.

Reacting to the dissonant data discovered by a team that did not implement methodological precautions, experts used all the hypothesized belief-system defenses, including challenging the authenticity of the archival documents, the representativeness of the documents, and the competence and the motives of the unnamed investigators. The open-ended data underscore this point. The same tactics of data neutralization were almost four times as likely to appear in spontaneous comments on dissonant than on consonant evidence (62 percent of the thought protocols produced by experts confronting dissonant data contained at least one evidence-neutralization technique versus 16 percent of the protocols produced by experts confronting consonant data, with the magnitude of the double standard about equal for experts on opposite sides of the median split of the ideology scale). When we create a composite belief-system-defense scale (the tendency to endorse all three tactics of data neutralization), the scale consistently predicts rejection of the conclusions that the investigators want to draw from their "discovery" (correlations ranging from .44 to .57 across historical periods).

It is tempting to see evidence here of hopeless closed-mindedness. Experts were far more responsive to the manipulation of empirical findings than to that of procedural quality, ignoring the latter information altogether when the data reinforced their ideological preconceptions and giving only some, and then rather grudging consideration to high-quality data that challenged their preconceptions. Before issuing a blanket condemnation, we should, however, consider three qualifications. First, the results do not show that experts ignore contradictory evidence. Some experts manifestly did change their minds in response to high-quality dissonant evidence. Second, the greater effect sizes for "empirical findings" than for "research procedures" might merely reflect that we manipulated evidence in a more compelling fashion than procedures. Comparisons of effect sizes across independent variables are notoriously problematic in the absence of a common metric. Third, the data do not demonstrate that experts are too slow to

Table 2. Average Expert Reactions to Dissonant and Consonant Evidence of Uncertain or High Quality Bearing on Three Controversial Counterfactuals

		LOW-QUALITY EVIDENCE			
High Feasibility of Counterfactuals		Impugn Motives of Investigators	Question Authenticity of Documents	Interpretation of Text	Overall Credibility
Purging Stalin	LIBERALS	3.0 (n=10)	2.9 (n=10)	3.8 (n=10)	7.1
	CONSERVATIVES	7.1 (n=9)	6.9 (n=8)	7.2 (n=8)	3.2
Ending Cold War in 1950's	LIBERALS	2.8 (n=10)	3.5 (n=10)	3.7 (n=10)	7.2
	CONSERVATIVES	7.0 (n=9)	6.6 (n=8)	6.8 (n=9)	2.9
Confrontational Soviet Response to Reagan	LIBERALS	3.4 (n=10)	3.9 (n=8)	3.1 (n=8)	6.9
	CONSERVATIVES	6.7 (n=9)	7.0 (n=9)	7.4 (n=9)	3.4
		HIGH-QUALITY EVIDENCE			
Purging Stalin	LIBERALS	2.8 (n=10)	2.5 (n=10)	3.2 (n=10)	7.3
	CONSERVATIVES	6.1 (n=9)	6.4 (n=9)	5.9 (n=9)	4.4
Ending Cold War in 1950's	LIBERALS	3.0 (n=10)	2.6 (n=10)	3.1 (n=10)	7.0
	CONSERVATIVES	5.7 (n=9)	6.0 (n=9)	6.4 (n=9)	4.7
Confrontational Soviet Response to Reagan	LIBERALS	3.1 (n=10)	2.8 (n=10)	2.7 (n=10)	7.0
	CONSERVATIVES	6.0 (n=9)	5.5 (n=9)	7.1 (n=9)	4.3

(continued on next page)

Table 2. Average Expert Reactions to Dissonant and Consonant Evidence of Uncertain or High Quality Bearing on Three Controversial Counterfactuals *(continued)*

Low Feasibility of Counterfactuals		Impugn Motives of Investigators	Question Authenticity of Documents	Interpretation of Text	Overall Credibility
LOW-QUALITY EVIDENCE					
Purging Stalin	LIBERALS	6.6 (n=10)	6.2 (n=10)	6.8 (n=10)	3.5
	CONSERVATIVES	3.1 (n=10)	3.5 (n=10)	4.0 (n=10)	6.9
Ending Cold War in 1950's	LIBERALS	5.7 (n=10)	5.9 (n=10)	6.5 (n=10)	4.0
	CONSERVATIVES	4.1 (n=10)	4.2 (n=10)	3.3 (n=10)	7.2
Confrontational Soviet Response to Reagan	LIBERALS	6.5 (n=10)	7.0 (n=10)	6.8 (n=10)	3.6
	CONSERVATIVES	2.6 (n=10)	3.4 (n=10)	3.9 (n=10)	6.7
HIGH-QUALITY EVIDENCE					
Purging Stalin	LIBERALS	5.2 (n=9)	5.4 (n=9)	6.1 (n=9)	4.2
	CONSERVATIVES	2.9 (n=8)	3.6 (n=8)	3.7 (n=8)	7.1
Ending Cold War in 1950's	LIBERALS	5.5 (n=9)	5.2 (n=9)	5.8 (n=9)	4.4
	CONSERVATIVES	4.0 (n=8)	3.8 (n=8)	3.1 (n=8)	7.0
Confrontational Soviet Response to Reagan	LIBERALS	6.1 (n=9)	6.8 (n=9)	6.2 (n=9)	4.3
	CONSERVATIVES	2.5 (n=8)	3.5 (n=8)	3.7 (n=8)	6.9

Higher scores indicate stronger endorsement of the hypothesized defensive cognitions.

accept dissonant evidence. It may be prudent to ask sharp questions of unexpected results.

The key normative point, however, remains the pervasiveness of double standards: experts switched on the high-intensity search light of skepticism only for dissonant results. Whether we trace the problem to excessive skepticism toward dissonant data or insufficient skepticism toward consonant data, counterfactual beliefs often appear self-perpetuating, effectively insulated from disconfirming evidence by a protective belt of defensive maneuvers and further reinforced by an understandable disinclination to attribute confirming evidence to either methodological sloppiness or to partisan bias. Tellingly, no one spontaneously said "the methodological errors just happened to break in my direction this time."

3. Theory-Driven Defenses of Conditional Forecasts: Studies VI–XII

Assessing the "truth value" of conditional forecasts initially looks more promising than assessing that of historical counterfactuals. When I "retrodict" that "if X had occurred, Y would have," there is no way of decisively refuting me. But when I predict that "if X occurs, then Y will occur," and X indisputably occurs and Y equally indisputably does not, I am obliged to qualify or abandon my original claim. Closer analysis reveals, however, that judging conditional forecasts in world politics is typically far more controversial than this atypical example suggests. Experts have at least five logically defensible strategies for protecting conditional forecasts that run aground troublesome evidence:

(1) The antecedent was never adequately satisfied—in which case, the conditional forecast becomes an historical counterfactual with the passage of time. Thus, experts might insist that "if we had properly implemented deterrence or reassurance, we could have averted war" or "if real shock therapy had been practiced, we could have averted this nasty bout of hyperinflation";

(2) Although the specified antecedent was satisfied, key background conditions (covered by the ubiquitous ceteris paribus clause) took on unexpected values, thereby short-circuiting the otherwise reliably deterministic connection between cause and effect. Experts might defiantly declare that rapid privatization of state industries would have led to the predicted surge in economic growth but only if the government had pursued prudent monetary policies;

(3) Although the predicted outcome did not occur, it "almost occurred" and would have but for some inherently unpredictable exogenous shock. Examples of such "close-call counterfactuals" (Kahneman and Varey 1990) include "the hardliners almost overthrew Gorbachev" and "the EU almost disintegrated during the currency crises of 1992");

(4) Although the predicted outcome has not yet occurred, it eventually will and we just need to be more patient (hardline communists may yet prevail in Moscow and the EU still might fall apart);

(5) 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 (Almond and Genco 1977; Jervis, 1992). As Henry Kissinger wryly conceded to Daniel Moynihan after the fragmentation of the Soviet Union, "your crystal ball worked better than mine" (Moynihan 1993, 23). On close inspection, this concession concedes nothing.

The data will show that expert observers of world politics draw on all five strategies of minimizing the conceptual significance of unexpected events. Tempting though it is to dismiss such intellectual maneuvering as transparently defensive post hocery, it would be wrong to issue automatic indictments for the judgmental bias of belief perseverance. Each defense highlights a potentially valid objection to viewing disconfirmation of the conditional forecast as disconfirmation of the underlying theory. Objections of this sort were indeed partly responsible for the abandonment of simple (Popperian) falsificationism in favor of more complex (Lakatosian) variants of falsificationism within the philosophy of science (Suppe 1973). For our purposes, it is not necessary to stake out a detailed position on the merits of specific variants of falsificationism. It is sufficient to specify a straightforward procedural test of bias that, if failed, would convince even the most forgiving falsificationist that something is awry. This de minimis test poses the question: When evidence arises bearing on a conditional forecast, do judges who "got it wrong" display much greater interest than judges who "got it right" in questioning whether the antecedent was satisfied, in generating close-call counterfactuals, and in otherwise challenging the probity of the exercise? If so, we still cannot determine who is biased (incorrect forecasters may be too quick to complain about the test or correct forecasters may be too quick to accept it) but we can say that some bias—in the form of theory-driven selectivity of information processing—does exist.

But the normative analysis need not stop here. It is possible in principle, and sometimes in practice, to construct a Bayesian benchmark for gauging experts' willingness to change their prior beliefs in accord with diagnosticity ratios that can be constructed from conditional-probability judgments that these same experts made of the conceptual relevance of potential outcomes when they advanced their original forecasts. Key questions become: Do experts who "get it wrong" retain more confidence in their prior understanding

of the underlying forces at work than they should have in light of earlier assertions they made about the probability of particular events, assuming the correctness of either their own or alternative interpretations of those underlying forces? Do experts who "get it right" take these confirmations too seriously, inflating their confidence in underlying forces at work even more than they should? And what cognitive mechanisms generate these effects? Is Bayesian underadjustment of "priors" among inaccurate experts, for example, mediated by reliance on the five belief-system defenses discussed earlier? These experts may see little reason to adjust their prior beliefs in accord with diagnosticity ratios provided long ago when they now believe that the ceteris paribus clause was not satisfied, the predicted consequences almost happened, and prediction is inherently impossible. By contrast, experts who made correct forecasts may "overadjust." They may be proud of their accomplishment and disinclined to question it too closely by raising awkward questions about whether their forecasts turned out right for the wrong reasons.

3.1 Method

Over the last twelve years, Tetlock (1992b, 1998) has been collecting experts' predictions of a wide array of political, economic, and military outcomes. The total sample of expert participants now exceeds 200 and the total number of predictions exceeds 5,000. The necessary time has not elapsed for testing many of these predictions (which often reach 10 and occasionally 25 years into the future), but it is possible to examine a subset for which adequate outcome evidence is available.

This article will consider the predictions that *only relevant area-study specialists* offered for the 5-year futures of the Soviet Union in 1988 (n=38), of South Africa in 1989 (n=26), of Kazakhstan in 1992 (n=19), of the European Monetary Union in 1991 (n=29), and of Canada in 1992 (n=29). It will also examine the predictions for the U.S. presidential race of 1992 in a 4-month time frame (n=34) and for the Persian Gulf crisis/war of 1990–1991 in a 1-year frame (n = 24). All participants had received some graduate training in social science or history, specialized in the region under examination, and earned their livelihoods either as advanced graduate students and professors in universities, policy analysts in think-tanks, intelligence analysts in government service, or journalists in the employ of the mass media.

The appendix presents in detail the instructions, assurances, and questions given to the experts. In essence, experts were asked to rate the likelihood that: (1) their understanding of the underlying forces shaping events in [x] was correct; (2) the most influential alternative interpretation of the underlying forces was correct; (3) various possible futures for [x] would occur assuming that: (a) their own understanding of underlying forces shaping events is correct; (b) the most influential alternative interpretation of the un-

derlying forces shaping events is correct. Respondents were given detailed guidance on how to use the subjective-probability scales that ranged from 0 to 1.0. The "possible futures" were designed to be logically exclusive and exhaustive. For example, in the Soviet case, the scenarios included a strengthening, a reduction, or no change in communist party control (for the other cases, see the appendix). After the specified forecasting interval had elapsed, 78 percent of the original forecasters were successfully contacted and questioned again. After exploring experts' ability to recall their original answers, experts were reminded of their original forecasts and confidence estimates. Experts rated their agreement with nine propositions that could theoretically either cushion the disappointment of "disconfirmation" or deflate the euphoria of "confirmation." Experts also answered "retrospective-probability" questions that permit some assessment of the degree to which experts updated their prior probabilities in an approximately Bayesian fashion (although this was done in only four of the seven forecasting domains examined here).

3.2 Results

Across all seven scenarios, experts were only slightly more accurate than one would expect from chance. Almost as many experts as not thought that the Soviet Communist Party would remain firmly in the saddle of power in 1993, that Canada was doomed by 1997, that neo-fascism would prevail in Pretoria by 1994, that the EMU would collapse by 1996, that Bush would be reelected in 1992, and that the Persian Gulf crisis would be resolved peacefully. Moreover, although experts only sporadically exceeded chance predictive accuracy, they regularly assigned subjective probabilities that exceeded the scaling anchors for "just guessing." In this sense, the results replicate the well established overconfidence effect for difficult items (Dawes, 1998). Most respondents thought they knew more than they did. Moreover, the margins of error were larger than those customarily observed in confidence-calibration research. Across all seven predictions, experts who assigned confidence estimates of 80 percent or higher were correct only 45 percent of the time.

Our principal interest here was, however, in neither forecasting accuracy nor confidence calibration, but rather in reactions to the apparent confirmation or disconfirmation of subjective forecasts. Not surprisingly, experts whose most likely scenarios did materialize credited their accuracy to their sound reading of the "basic forces" at play. Across issue domains they assigned average ratings between 6.6 and 7.3 on a nine-point scale where 9 indicates maximum confidence. More surprisingly, experts whose most likely scenarios did not materialize were almost as likely to believe that their reading of the political situation was fundamentally sound. They assigned average ratings from 6.3 to 7.1.

How did experts who "got it wrong" convince themselves that they were basically right? The cognitive-consistency hypothesis was that these forecasters preserved confidence in their world view by invoking various belief-system defenses. Table 3 summarizes endorsements of five key defenses: arguing that the antecedent and ceteris paribus clauses underpinning the original forecasts were not satisfied, invoking close-call counterfactuals, claiming that the predicted event might still occur, and minimizing the conceptual significance of forecasting exercises. In twenty-six of thirty-five cases, t-tests reveal that experts whose most likely scenarios did not occur showed significantly more enthusiasm ($p < .05$) for "defensive" cognitions than experts whose most likely scenarios did occur.

One of the most popular defenses—reflected both in the rating-scale data and in the spontaneous comments—was the close-call counterfactual claim that the predicted outcome "almost occurred." This defense surfaced in all seven forecasting domains:

(1) in contrast to observers who foresaw reduced party control, Sovietologists who in 1988 predicted continued or greater communist party control over the next 5 years were more prone to claim that hardliners almost succeeded (e.g., in the coup attempt of August 1991) and might well have succeeded had it not been for their own monumental ineptitude and the courage of political and military leaders who resisted the coup;

(2) observers of the U.S. scene who expected Bush to be reelected found it easier to imagine a world in which Clinton never became president than did those who foresaw a Clinton win. Their close-call counterfactuals posited, for example, a more compliant Federal Reserve Board (cutting interest rates earlier in 1991) and a deft campaign of negative advertising aimed at Clinton's character;

(3) observers of South Africa who expected continued or even increasingly oppressive white-minority rule from 1989 to 1994 were especially likely to believe that were it not for the coincidental conjunction of two key individuals—de Klerk and Mandela—in key leadership roles, South Africa could easily have gone down the path of increasing repression, racial polarization, and violence;

(4) observers who viewed Kazakhstan as "several Yugoslavias waiting to erupt into interethnic violence" tended to attribute the nonoccurrence to the shrewdness of the Kazakh leadership as well as to the lack of interest of current Russian leaders in playing the "ethnic card";

(5) experts who expected the European Monetary Union to collapse argued that the event almost happened in the wake of the currency crises of 1992 and indeed would have occurred had it not been for the principled determination (or perhaps obstinacy) of key politicians. Given the deep conflict of interest between states that have "solid fundamentals" and those that

Table 3. Average Reactions of Experts to Confirmation and Disconfirmation of Their Conditional Forecasts

Belief-System Defenses:		Close-call Counter-factuals	Ceteris Paribus Did Not Hold	Antecedent Did Not Hold	Just Unlucky About Timing	Dismiss Forecasting Exercises in General
Predicting Future of:	Status of Forecast					
Soviet Union	Inaccurate	7.0*	7.1*	6.8*	6.4	7.3*
	Accurate	4.1	3.9	3.6	5.0	3.1
South Africa	Inaccurate	7.1*	7.0*	7.3*	3.7	7.1*
	Accurate	4.5	3.5	3.3	4.0	4.8
EMU	Inaccurate	7.2*	5.9	6.2	7.8*	7.0*
	Accurate	5.1	4.6	4.9	3.8	4.3
Canada	Inaccurate	7.6	6.8*	6.5*	8.0*	7.2*
	Accurate	6.8	3.7	4.2	4.4	4.5
Kazakhstan	Inaccurate	6.5*	7.2*	6.9*	7.3*	6.8*
	Accurate	2.8	2.9	3.5	3.6	5.0
Persian Gulf Crisis	Inaccurate	7.0*	6.6	6.4	4.4	6.9*
	Accurate	4.1	4.8	4.5	4.2	4.1
1992 Presidential Election	Inaccurate	7.5*	6.7*	6.6*	3.5	6.9*
	Accurate	5.8	3.2	3.1	3.6	5.2

* indicates significantly stronger endorsements of the belief-system defense among inaccurate forecasters ($p < .05$).

"regularly resort to creative accounting to make their budget deficits appear smaller" and given nationalist resentment of a single European currency, these experts thought it a "minor miracle" that European leaders largely remained in 1997 committed to monetary union, albeit on a loophole-riddled schedule;

(6) experts who anticipated the secession of Quebec pointed to how nearly the second separatist referendum passed (if a fraction of a percentage point of the electorate had voted "oui" rather than "non",. . .) and to how a more effectively managed campaign could have easily won a larger fraction of the swing vote (if a more savvy and charismatic politician, say Bouchard rather than Parizeau, had spearheaded the cause,. . .);

(7) experts who expected Saddam Hussein to recognize during the Gulf Crisis that the balance of power was tipping irreversibly against him and to withdraw from Kuwait were more likely to claim that Saddam would have acted as they predicted had he only understood as clearly as did the experts the true configuration of forces. They were also inclined to attribute the tragedy to Saddam's pathological personality which predisposed him to act far more recklessly than most heads of state. One expert complained: "How can anyone know what sets him off? Perhaps he convinced himself that it was better to be a rooster for one day than a chicken for all eternity. But judging from his record, he could have latched on to another proverb and convinced himself that it was better—like Saladin—to retreat and return to fight another day."

Experts whose most likely scenarios did not materialize also argued that the antecedent and ceteris paribus clauses underlying the original forecasts had not been satisfied because a qualitatively new array of "fundamental forces" had come into play. For example, some observers of the Persian Gulf crisis who predicted that Saddam would do the "rational thing" and fracture the U.S.-led alliance by preemptively withdrawing from Kuwait argued that this course of action had ceased to be a viable face-saving option for the Iraqi leadership because the U.S. had calculatingly blocked it. The meaning of rationality had been transformed. The geopolitical definition that inspired the original forecast had been superseded by a cultural and domestic political one.

Relatively inaccurate forecasters also insisted that they had just been "unlucky" on timing. This defiant defense was especially popular among those who predicted the demise of Canada (the Parti Québecois will try again and prevail on its third attempt), Kazakhstan (demagogues on both sides of the border with Russia will seize on the opportunities for ethnic mobilization that Kazakhstan presents), and the European Monetary Union (the divergent interests of the prospective members will trigger crises that even determined leadership cannot resolve). In effect, the experts declared

that we may have been wrong within this arbitrary time frame but we shall eventually be vindicated.

Finally, relatively inaccurate forecasters were more dismissive of forecasting exercises than were experts who got it right. The two intercorrelated measures of this strategy ($r = 0.46$) are of special interest because they were assessed both before experts made their predictions and after they learned of the accuracy of their predictions. Inaccurate forecasters shifted significantly toward the view that: (1) politics is inherently indeterminate; (2) forecasting exercises are deeply misleading because they assign too much credit to "winners" (who may be merely lucky) and too much blame to "losers" (who may be merely unlucky). By contrast, accurate forecasters shifted, albeit nonsignificantly, in the opposite direction on these indicators.

It is striking that forecasters who had greater reason to be surprised by subsequent events managed to retain nearly as much confidence in the fundamental soundness of their judgments of political causality as forecasters who had less reason to be surprised. It is also striking that relatively inaccurate experts were more likely than their accurate colleagues to endorse propositions that had the logical effect of either insulating conditional forecasts from troublesome evidence or minimizing the significance of the evidence. From this pattern, it is tempting to conclude that we have identified the cognitive mechanisms by which experts sustain confidence in the wake of unexpected events: a five-tiered defensive perimeter of abandoning the antecedent, challenging ceteris paribus, close-call counterfactualizing, acknowledging only an error of timing, and dismissing forecasting exercises in general by stressing the indeterminate character of politics.

But temptation is not proof. Our thus far circumstantial case can be substantially strengthened by articulating a more detailed model of the psychologic of belief-system defense and subjecting that model to additional tests:

(1) if belief-system defenses are activated by big surprises, experts should most strongly endorse "defensive" cognitions to the degree that, at the original forecast, they expressed confidence in scenarios that did not materialize. The psycho-logic is straightforward: the greater the confidence in the original forecast, the more threatening the apparent disconfirmation to experts' claims to expertise, and the more motivated experts will be to neutralize the troublesome evidence. All else equal, an expert who in 1988 was 90 percent confident that communist hardliners would reassert control between 1988 and 1993 should be more unsettled by intervening events than an expert who attached only "guessing" confidence to the same forecast. To test this prediction, we created a composite belief-system-defense index by summing the five indicators in Table 3. The predicted pattern emerged. Among relatively inaccurate forecasters, the correlations between ex ante confidence (original forecast) and scores on the composite belief-system-

defense index are always positive, ranging from 0.28 to 0.44 across domains; among relatively accurate forecasters, the same correlations hover between –.07 and +0.11;

(2) if belief system defenses cushion the blow of unexpected events, then experts whose most likely scenarios do not materialize but who endorse "defensive" cognitions should be better able to retain confidence in their original forecasts after they learn of what happened (ex post confidence). But there should be no such correlation among experts whose conditional forecasts were borne out and who should therefore not have perceived any threat to the core tenets of their belief systems. As predicted, among relatively inaccurate forecasters, the defensiveness index is correlated with ex post confidence across domains (correlations ranging from .21 to 0.41). By contrast, among relatively accurate forecasters, there is almost no relationship between defensiveness and ex post confidence (correlations between –.02 and .11);

(3) if belief-system defenses permit experts to preserve confidence in basic beliefs when unexpected outcomes occur, the correlations between ex ante confidence in the original forecasts and ex post confidence in those same forecasts should be positive for both relatively accurate forecasters and their less accurate colleagues who deployed "defensive" cognitions to protect their belief systems. The correlations should however be much lower for experts who did not "close" the gaps between conditional forecasts and political outcomes and who thus "hemorrhaged" confidence. With two noteworthy exceptions—Persian Gulf and E.M.U.—which yielded null results, the data fit this pattern. For relatively accurate experts, correlations between ex ante and ex post confidence in the other five domains were consistently significant, ranging between 0.25 and 0.36; for relatively inaccurate experts who defended their belief systems, the correlations were comparable, between 0.22 and 0.43. But these correlations fell close to zero, ranging from –0.05 to 0.14, for experts whose most likely scenarios did not occur but who failed to defend their belief systems (arguably due to the decline of ex post confidence in this latter group);

(4) if the five belief-system defenses do indeed serve the same underlying cognitive function, then these defensive strategies should be positively intercorrelated but only up to a point. On the one hand, a common cause—concern for defending the correctness of one's understanding of underlying forces—should produce shared variance. Consistent with this hypothesis, the average intercorrelation among the five "defensive" cognitions is 0.29. On the other hand, correlations should not edge much higher for two reasons. First, certain defenses become implausible, and therefore useless for preserving professional self-esteem, in certain domains. No one offers the "off-on-timing" defense for single-occurrence events like the 1991 Gulf War

or the 1992 U.S. presidential election. Second, some experts may conclude that a single defense is adequate. If invoking a close-call counterfactual is subjectively sufficient to eliminate the threat to one's belief system, why bother generating additional defenses that will only strike observers as, frankly, defensive?

4. General Discussion

Taken together, the three sets of studies underscore how easy it is even for sophisticated professionals to slip into borderline tautological patterns of thinking about complex path-dependent systems that unfold once and only once. The risk of circularity is particularly pronounced when we examine reasoning about ideologically charged historical counterfactuals. Although it is understandable that experts in Studies I and II relied heavily on their preconceptions to fill in what would have happened in the hypothetical control conditions of history (what else could they do?), it is much more difficult to justify the reluctance of experts in Studies III–V to reconsider their theory-driven beliefs about counterfactual worlds when confronted by potential evidence from the actual world that undercut those beliefs (a reluctance that is especially difficult in light of experts' willingness to accept exactly the same evidence when it tipped the scales in a congenial direction). Circularity also seems to be a serious problem for experts' reasoning about conditional forecasts. In Studies VI–XII, experts resisted acknowledging apparent disconfirmation by resorting to a host of strategies that, ironically, had the net effect of transforming past-due conditional forecasts into historical counterfactuals of the form "if the antecedent condition or ceteris paribus clause had been satisfied, the predicted event would have happened," or "the predicted event almost occurred and would have but for an unforeseeable accident of history." Moreover, these belief-system defenses worked: experts who got it wrong were often as confident as experts who got it right in their readings of the underlying forces that shaped events. Given our habits of thought and the structure of the situation, it is tempting to conclude that it is practically impossible for experts to learn anything from history that they were not already cognitively predisposed to learn.

How far should we take this pessimistic thesis? Are people irredeemably theory-driven thinkers who automatically assimilate evidence into extant knowledge structures, pausing only briefly for disconfirmed predictions to generate a defensive counterfactual? Or is there some way to reconcile the data with a more rationalistic, perhaps Bayesian, account? The data initially look unpromising for defenders of human rationality. Table 4 suggests that if we are determined to hold the hypothesis that most experts are good intuitive Bayesians, we must also defend the defensive strategies these experts use for downplaying the diagnosticity of dissonant data as well as the offensive

Table 4. Subjective Probabilities That Experts Assigned Their Understanding of Underlying Forces at Beginning and End of Forecasting Periods

Predicting Future of:	Status of Forecast	Actual Prior Probability	Actual Posterior Probability	Bayesian Predicted Posterior Probability
Soviet Union	Inaccurate	0.74	0.70	0.49[b]
	Accurate	0.69	0.83[a]	0.80
South Africa	Inaccurate	0.72	0.69	0.42[b]
	Accurate	0.70	0.77	0.82
EMU	Inaccurate	0.66	0.68	0.45[b]
	Accurate	0.71	0.78	0.85
Canada	Inaccurate	0.65	0.67	0.39[b]
	Accurate	0.68	0.81[a]	0.79

(a) indicates significant shift in subjective probability assigned to one's understanding of underlying forces at beginning and end of forecasting (prior versus posterior).

(b) indicates a significant deviation of expert-assigned posterior probabilities from the posterior probabilites that Bayes Theorem stipulates experts should endorse if they were correctly multiplying the prior odds by the diagnosticity ratios provided at original forecast.

strategies they use for inflating the diagnosticity of consonant data. Across all four domains with measures of prior probabilities, diagnosticity ratios at the original forecasts, and posterior probabilities at the follow-up session, experts whose most likely scenarios materialized increased their estimates of the likelihood of their prior understanding of underlying forces being correct whereas experts whose most likely scenarios failed to materialize showed no inclination to decrease their estimates of the likelihood of their prior understanding of underlying forces being correct. The data look especially bad for a descriptive Bayesian model of expert judgment when we factor in the diagnosticity ratios constructed by dividing: (a) experts' ex ante assessments of the likelihoods of different scenarios given the hypothesis that experts' understanding of underlying forces was correct by (b) experts' ex ante assessments of the likelihood of the same scenarios given the hypothesis that the most influential alternative view was correct. When we multiply the prior odds by these diagnosticity ratios for each respondent's forecasts, Bayes' Theorem tells us how much experts *should* change their minds in the correctness of their initial hypotheses. Here we discover that relatively inaccurate forecasters do not shift their posterior probabilities nearly as much as Bayes' theorem stipulates. These findings replicate the well known "conservatism bias" in probabilistic reasoning: people do not change their minds as much as a good Bayesian would.

Even these data are not, however, decisive. In the intervening (usually five year) period, experts may have quite sensibly changed their minds about the diagnosticity of various events vis-a-vis various hypotheses. Indeed, one can make a good case that each of the belief-system defenses documented here invokes a persuasive reason why relatively inaccurate experts should not abandon prior probability assessments that were conditional on the soundness of the experts' original understanding of underlying forces. Why change one's mind about the validity of one's pre-forecast conceptual framework when the necessary antecedent conditions for applying that framework were never fulfilled, exogenous shocks vitiated the ceteris paribus requirement for all fair tests of hypotheses, the predicted event almost occurred despite all these obstacles and still might, and prediction exercises are so riddled with indeterminacy as to be meaningless? These are all plausible reasons for supposing that, whatever diagnosticity ratios experts implicitly endorsed years ago, the new diagnosticity ratios assign subjective probabilities to P (D/H) and P (D/~H) that justify a refusal to change one's mind (in other words, the ratios converge on 1). This analysis, in turn, suggests that the focus on whether experts got their Bayesian arithmetic right is misplaced. The more fundamental question concerns the defensibility of each belief-system defense. Let's consider each in turn:

(1) *Nonfulfillment of antecedents and ceteris paribus clauses.* There are sound philosophy-of-science grounds for refusing to abandon a theory until one is convinced that the antecedent conditions for activating the relevant covering laws have been fulfilled and the ceteris paribus clause has been satisfied (Hempel 1965). Experts often had reason for suspecting that the "fundamental forces" on which forecasts were initially predicated were not the same forces shaping events at the end of the forecasting period. Subsequent events had transformed the original meanings of key analytical concepts—rationality, economic determinism, the perceived interests of former apparatchiks, or the white business elite—underlying forecasts. And background conditions—too numerous to specify in advance—occasionally took on unexpectedly extreme values that routed events down "unforeseeable" paths;

(2) *Close-call counterfactuals.* Political outcomes often appear to be under the control of complex conjunctions of stochastic processes that could easily have converged on different values and produced different effects. One's reaction to this defense hinges on one's metaphysical outlook. LaPlacean determinists will reject it out of hand but if one accepts Gould's (1995) provocative thought experiment in which, holding initial conditions constant, repeatedly rerunning the tape of evolutionary history yields thousands of distinctive outcomes (with intelligent life one of the least likely), it is reasonable to talk about alternative histories "almost happening" and

even to distinguish alternative histories on the basis of how "close" they came to happening. It is certainly not far-fetched to claim that the Parti Québecois almost won the second secessionist referendum, and it arguably is not far-fetched to claim that hardline communists nearly reclaimed power in the late Gorbachev period. Indeed, if we interpret the subjective-probability estimates attached to forecasts literally as "frequentist" claims about the distribution of possible worlds in repeated-simulation reruns of history, experts who assigned 80 percent confidence to "incorrect" forecasts may quite literally have been correct. Strange though it sounds, in the ontology of Lewis' (1973) modal logic, it makes sense to say that in 80 percent of the Soviet Unions that were possible in 1988, the predicted outcome occurred. We just happened to wind up in that unlikely subset of possible worlds in which communist control collapsed. Of course, this defense wears thin with repeated use: People will tire of hearing that the world they happen to inhabit is vanishingly improbable;

(3) *Off-on-timing*. This defense is easy to ridicule. It reminds people of political jokes such as the probably apocryphal account of a Trotskyite rally at which the speaker enthusiastically declares Leon Trotsky to be a remarkably far-sighted man and to offer as proof that "of all the predictions Trotsky has made not one has yet come true." Nonetheless, this defense merits a serious hearing. The domains in which the defense was most frequently invoked—Russia, the E.M.U., and Canada—were those in which experts were especially unsure, ex ante, of the temporal boundaries to place on their forecasts. Even now, there is residual disagreement over whether the Russian communist party or some other authoritarian force will yet triumph and over whether Canada and the E.M.U. will dissolve. It seems silly to classify these forecasts as categorically wrong if they turn out to be approximately correct, albeit a bit tardy;

(4) *Conceptual Significance of Prediction*. This defense takes the indeterminacy argument to its ultimate extension. Rather than invoking indeterminacy in piece-meal fashion (a close-call counterfactual here or there), experts proclaim politics to be hopelessly "cloud-like," disparage those who look for lawful regularities as subscribing to a naive "clock-like" perspective, and dismiss forecasting exercises as trivial at best (a game) or deeply misleading at worst (the winners deserve no more credit for good judgment than does the roulette player who repeatedly bets on his favorite nephew's birthdate and occasionally wins). These experts are embracing a widely held philosophical position with radical implications for the notion of good judgment. There should be as little consistency in who wins political forecasting tournaments as there apparently is in forecasting tournaments for financial markets where random-walk models are widely held to hold (Malkiel 1990).

Widespread reliance on these belief-system defenses is not only compatible with various conceptions of rationality; it is also compatible with several conceptions of learning from history. Indeed, the data shed light on when learning is likely and on the forms that it will take. It will often be stimulated by failure (Argyris and Schoen 1996; Levy 1994) and constrained by the cognitive and organizational principles of least resistance to take incremental forms (Nye 1988; Stein 1994; Tetlock 1991). Most experts initially respond to unexpected events by tinkering with peripheral cognitions that require minimal modification of their belief systems (cf. Axelrod 1976; Holsti 1967; Khong, 1991). Rather than conceding on "fundamentals," they argue that some new unforeseeable causal force emerged, thus nullifying the ceteris paribus clause, or that the predicted outcome "almost occurred." The cumulative effect of such tinkering should not be understated. Whenever experts argue that their predicted future would have occurred but for some historical accident or previously unacknowledged moderator variable (often of a micro-nature such as leadership), they have learned in a significant sense. Their belief systems now assign a more important role to chance and recognize causality to have been more complex than previously acknowledged. From this standpoint, the more defensive counterfactuals one generates, the greater the likelihood that one is learning both in the cognitive-structural meaning of the term (evolving differentiated and integrated knowledge structures for processing future evidence) and in the efficiency sense of the term (acquiring the ability either to make more accurate forecasts or to attach more realistic confidence estimates to one's forecasts).

This argument suggests a curious conclusion: the more often experts are wrong, the wiser they become as they acknowledge more qualifications on when their expectations hold. The real threat to good judgment lies in the hubris that builds up from a succession of predictions that serendipitously turn out right. We thus have a substantive cognitive argument—in addition to the formal statistical argument invoking regression toward the mean—for expecting the poor predictive performers of yesterday to catch up to the superstar predictors of today.

How should we balance these conflicting arguments bearing on the "rationality" of professional observers of world politics? Here it is useful to recall the Faustian compromises underlying our research designs. The goal was to study how experts think about problems that arise naturally in their professional lives: extracting lessons from history, changing one's mind about historical causation in response to evidence, and responding to the apparent confirmation or disconfirmation of conditional forecasts. In making this choice, we sacrificed the precise normative benchmarks for error that laboratory researchers enjoy. Respondents did not randomly draw red or

blue marbles from an urn and then judge the likelihood of various distributions of colors—a task with a clear right or wrong Bayesian answer. Rather, they dealt with subtly interrelated and constantly evolving path-dependent sequences of arguably unique events (certainly, events with difficult-to-define base rates). As a result, even with ex ante diagnosticity ratios in hand, it is extraordinarily difficult to single out any given individual as biased.

The case for deviations from rationality can, however, be made in additional ways. Most important, it can be pieced together not from individual judgments but rather from aggregations of judgments that bring double standards into clear focus. Defined thusly, bias takes the form in Studies III–V of looking for methodological flaws only in evidence that challenges favorite historical counterfactuals. In Studies VI–XII, bias takes several forms, most notably, the strong interest of less accurate forecasters in downplaying the relevance of subsequent events to their prior understanding of the basic forces at work and the much more subdued interest of the same forecasters in adjusting prior beliefs in a Bayesian fashion in accord with diagnosticity ratios constructed from conditional-probability judgments offered at the original forecasts. Characterizing these tendencies as defensive does not require granting the author any mysterious ontological insights into "how truly close" any given political entity came to being re-routed onto an alternative historical trajectory. Rather the case ultimately rests on epistemological intuitions about what constitutes fair play in the hypothetico-deductive game. Whether these intuitions take precise Bayesian or vague Lakatosian forms, the patterning of data is suggestive of both excessive conservatism in belief revision and of excessive reliance on a protective belt of auxiliary hypotheses that cushion implicit theories from refutation by severing logical links (~q therefore ~p).

There is also a second argument that experts strayed from rationality, however warily that concept must be defined in real-world contexts. The results on counterfactual reasoning in Studies III–V and on forecasting in Studies VI–XII replicate key findings derived from laboratory paradigms with precise normative benchmarks for defining error and bias. The skepticism that experts reserved for dissonant discoveries extends the work of Lord, Ross, and Lepper (1979) on theory-driven assessments of evidence as well as the work of Wilson et al. (1993) on selective standards of proof in scientific hypothesis testing (Wilson et al. 1993); the overconfidence experts displayed in their forecasts reaffirms a massive body of work on calibration of subjective probability estimates of knowledge (Fischhoff 1982); the selective activation of belief-system defenses dovetails nicely with the classic dissonance prediction that people would most need "defenses" when they appear to have been wrong about something in which they were originally quite confident (Festinger 1964); the generation of close-call counterfactuals

in response to unexpected events is consistent with experimental evidence on determinants of spontaneous counterfactual thinking (Kahneman and Miller, 1996); 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 (Einhorn and Hogarth 1981). In all five respects, the current results underscore the generalizability of laboratory-based demonstrations of bounded rationality in a more ecologically representative research design. The psychological findings 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).

Still, some observers may dismiss the results as but one more "debunking" of expertise in general or of the forecasting powers of social scientists in particular. It is easy to curse the epistemological darkness, harder to light a methodological candle. In closing, therefore, it seems appropriate to explore the prospects for improving expert judgment. One school of thought in cognitive psychology will be deeply skeptical that much can be done. In this view, the human mind has been designed by natural selection less for dispassionate balancing of evidence than it has been for rapidly reaching conclusions via the application of low-effort heuristics (cf. Arkes, 1991). But another more social psychological school of thought portrays the same patterns of expert judgment not as products of neurological hardwiring but rather as strategic adaptations to a professional culture in which one's reputation hinges on appearing approximately right most of the time and on never appearing clearly wrong. In this view, expert judgment takes the theory-driven forms it does because it has been conditioned more by the rhetorical demands of thrust and parry in an adversarial environment than by the logical demands of hypothetico-deductive method. This framework suggests grounds for optimism: if the dominant professional organizations shifted their policies toward rewarding self-critical confessions of error and encouraging rigorous scrutiny of belief-system defenses, experts would strategically adapt to this new accountability regime (cf. Tetlock, 1992a). At this juncture, we counsel an agnostic stance toward these two perspectives. The psychological literature is itself divided between those who trace judgmental biases to fundamental perceptual and associative properties of the human mind and those who favor motivated-cognition accounts that endow humans with greater flexibility in deciding how to decide.

Finally, it is important to balance psychological and political considerations in interpreting these data. Psychologists often assume that people simply seek to maximize accuracy (Fiske and Taylor, 1991). Inasmuch as the

goal is exclusively accuracy, people should try to bring their self-assessments of cognitive prowess into closer alignment with their modest predictive achievements. This might be done using standard "de-biasing" manipulations: memory aids for recalling ex ante states of uncertainty as well as judgment aids for stimulating awareness of counterarguments and for detecting double standards via perspective-taking thought experiments. But the qualification "inasmuch" is critical. Experts may value accuracy only conditionally. One ideological camp may deem it most important both to score hits (e.g., identify true aggressors) and to avoid misses (failing to identify true aggressors), hence an "irrationally" high tolerance for false alarms (calling status quo powers aggressive); another camp may have the opposite priorities. Whether we think observers of world politics are doing a good job hinges not just on a simple signal-detection-accuracy score (say, Hits minus False Alarms); it hinges on the political values we place on maximizing or minimizing the key logical components of accuracy.

Manuscript submitted 16 January 1998.
Final manuscript received 16 September 1998.

REFERENCES

Almond, Gabriel, and T. Genco. 1977. "Clouds, Clocks, and the Study of Politics." *World Politics* 3: 489–522.

Argyris, Christopher, and Donald A. Schon. 1996. *Organizational Learning II: Theory, Method, and Practice*. Reading, Mass.: Addison-Wesley Publishing Co.

Arkes, Hal. 1991. "Costs and Benefits of Judgment Errors: Implications for Debiasing." *Psychological Bulletin* 110:486–498.

Axelrod, Robert. 1976. *Structure of Design*. Boston, Mass.: Little, Brown.

Breslauer, George. 1996. "Counterfactual Reasoning in Western Studies of Soviet Politics and Foreign Relations." In *Counterfactual Thought Experiments in World Politics*, ed. Philip E. Tetlock and Aaron Belkin. Princeton: Princeton University Press.

Cohen, Stephen F., Alexander Rabinowitch, and Robert Sharlet. 1985. *The Soviet Union Since Stalin*. Bloomington: Indiana University Press.

Dawes, Robyn. 1998. "Judgment and Choice." In *Handbook of Social Psychology*, ed. Daniel Gilbert, Susan Fiske, and G. Lindzey. New York: McGraw Hill.

Einhorn, Hillel, and Robin Hogarth. 1981. "Behavioral Decision Theory: Processes of Judgment and Choice." *Annual Review of Psychology* 31:53–88.

Elster, Jon. 1978. *Logic and Society: Contradictions and Possible Worlds*. New York: Wiley.

Fearon, James. 1991. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43:474–484.

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

Fischhoff, Baruch. 1975. "Hindsight Is Not Equal to Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty." *Journal of Experimental Psychology* 104:288–299.

Fischhoff, Baruch. 1982. "Debiasing." In *Judgment Under Uncertainty*, ed. D. Kahneman, P. Slovic, and A. Tversky. Cambridge, England: Cambridge University Press.

Fiske, Susan, and Shelley Taylor. 1991. *Social Cognition.* Reading, Mass.: Addison-Wesley.
Garthoff, Raymond. 1994. *Detente and Confrontation: American-Soviet Relations from Nixon to Reagan.* Washington, D. C.: Brookings Institution.
Gould, Stephen Jay. 1995. *Dinosaur in a Haystack: Reflections in Natural History.* New York: Harmony Books.
Griffin, Dale, and Lee Ross. 1991. "Subjective Construal, Social Inference, and Human Misunderstanding." In *Advances in Experimental Social Psychology Volume 24*, ed. M. Zanna. New York: Academic Press.
Hempel, Carl. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: Free Press.
Hogarth, Robin, and William Goldstein. 1996. (eds.) *Judgment and Decision Making: An Interdisciplinary Reader.* Cambridge, England: Cambridge University Press.
Holsti, Ole. 1967. "Cognitive Dynamics and Images of the Enemy." In *Enemies of Politics*, ed. R. Fagan. Chicago: Rand McNally.
Jervis, Robert. 1976. *Perception and Misperception in International Politics.* Princeton: Princeton University Press.
Jervis, Robert. 1992. "The Future of International Politics: Will It Resemble the Past?" *International Security* 16:39–73.
Kagel, John, and Alvin Roth. 1995. *The Handbook of Experimental Economics.* Princeton: Princeton University Press.
Kahneman, Daniel, and Dale Miller. 1986. "Norm Theory: Comparing Reality to Its Alternatives." *Psychological Review* 93:136–153.
Kahneman, Daniel, and Carol Varey. 1990. "Propensities and Counterfactuals: The Loser That Almost Won." *Journal of Personality and Social Psychology* 59:1101–1110.
Kahneman, Daniel, Paul Slovic, and Amos Tversky, ed. 1982. *Judgment Under Uncertainty: Heursitics and Biases.* New York: Cambridge University Press.
Khong, Y. F. 1991. *Analogies at War.* Princeton: Princeton University Press.
Levy, Jack 1994. "Learning and Foreign Policy: Sweeping a Conceptual Minefield." *International Organization* 48:279–312.
Lewis, David K. 1973. *Counterfactuals.* Cambridge: Harvard University Press.
Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* 37: 2098–2109.
Malkiel, Burton. 1990. *A Random Walk down Wall Street.* New York: Norton.
McGuire, William J. 1969. "The Nature of Attitude and Attitude Change." In *Handbook of Social Psychology*, ed. E. Aronson and G. Lindzey. Reading, Mass.: Addison Wesley.
Moynihan, Daniel Patrick 1993. *Pandemonium.* New York: Oxford University Press.
Nye, J. 1988. "Nuclear Learning and U.S.-Soviet Security Regimes." *International Organizations* 41:121–166.
Pipes, Richard. 1993. *Russia Under the Bolshevik Regime.* New York: A. A. Knopf.
Roese, Neil, and James Olson. 1995. "Counterfactual Thinking: A Critical Overview." In *What Might Have Been: The Social Psychology of Counterfactual Thinking,* ed. Neil J. Roese and James Olson. Hillsdale: Erlbaum.
Stein, Janice. 1994. "Political Learning by Doing: Gorbachev as Uncommitted Thinker and Motivated Learner." *International Organization* 48:155–183.
Suppe, Frederick 1973. *The Structure of Scientific Theories.* Chicago: University of Chicago Press.
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, Philip E. 1991. "Learning in U.S. and Soviet Foreign Policy: In Search of an Elusive Concept." In *Learning in U.S. and Soviet Foreign Policy*, ed. George Breslauer and Philip Tetlock. Boulder: Westview.

Tetlock, Philip E. 1992a. "The Impact of Accountability on Judgment and Choice: Toward a Social Contingency Model." In *Advances in Experimental Social Psychology Volume 25*, ed. M. Zanna. New York: Academic Press.

Tetlock, Philip E. 1992b. "Good Judgment in World Politics: Three Psychological Perspectives." *Political Psychology* 13:517–540.

Tetlock, Philip E., and Aaron Belkin. 1996. *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*. Princeton: Princeton University Press.

Wilson, Timothy D., Bella M. DePaulo, D. G. Mook, and K. G. Klaaren. 1993. "Scientists' Evaluations of Research: The Biasing Effects of the Importance of the Topic." *Psychological Science* 4:322–325.