Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment

Philip E. Tetlock,[1,2] Christopher Karvetski,[3] Ville A. Satopää,[4] and Kevin Chen[1]

Futures and Foresight Science, in press.

Author affiliations. 1: University of Pennsylvania, Wharton; 2: Forecasting Research Institute; 3: Good Judgment Inc; 4: INSEAD

Author Acknowledgements. The original research in 1988 was supported by the MacArthur Foundation and assisted by the Committee for the Prevention of Nuclear War (National Academy of Sciences). The extension in 1997 was supported by the Mershon Center (The Ohio State University). The write-up and analyses in 2022-2023 were supported by Open Philanthropy, Founders Pledge and the John F. Templeton Foundation. We appreciate the early-stage advice and connections of Robert Jervis, Alexander George, Robert Powell, Ernst Haas, John Steinbruner, Terry Busch, Pete Lavoy, Kennette Benedict, and Richard Herrmann. We also acknowledge the prodding and support of Luke Muelhauser, without whom this article might never have been written.

Abstract

Skeptics see long-range geopolitical forecasting as quixotic. A more nuanced view is that although predictability tends to decline over time, its rate of descent is variable. The current study gives geopolitical forecasters a sporting chance by focusing on slow-motion variables with low base rates of change. Analyses of 5, 10 and 25-year cumulative-risk judgments made in 1988 and 1997 revealed: (a) specialists beat generalists at predicting nuclear proliferation but not shifting nation-state boundaries; (b) some counterfactual interventions—e.g., Iran gets the bomb before 2022—boosted experts' edge but others—e.g., nuclear war before 2022—eliminated it; (c) accuracy fell faster on topics where expertise conferred no edge in shorter-range forecasts. To accelerate scientific progress, we propose adversarial collaborations in which clashing schools of thought negotiate Bayesian reputational bets on divisive issues and use Lakatosian scorecards to incentivize the honoring of bets.

Long-Range Subjective-Probability Forecasts of Slow-Motion Variables in World Politics: Exploring Limits on Expert Judgment

Predictability varies across levels of analysis. Physicists tell us events on quantum scales are irreducibly probabilistic but events on cosmological scales unfold with deterministic regularity over billions of years (Deutsch, 1998). Between these end-points are countless outcomes unfolding on human time scales—economic growth, life spans, trading patterns—that are blends of deterministic and probabilistic processes. Many scholars argue long-range predictability must decline in this zone due to the compounding of noise over time (Kahneman et al., 2021)—and the rate of decline is often steep. Meteorologists tell us tiny input errors (butterfly effects) to computer simulations of weather systems cause chaotic perturbations within weeks (Gleick, 2008). Macro-economists warn of their models' limited power to anticipate booms and busts (Goettzman, Kim & Shiller, 2022). Indeed, some observers have concluded the combination of path-dependency and sensitive dependence on initial conditions makes long-range forecasting flat-out impossible (Gleick, 2008; McCloskey, 1991; Taleb & Blyth, 2013).

This article contrasts two ideal-type schools of thought—Radical Skepticism and Moderate Meliorism—that Tetlock (2005) pitted against each other in 1-to-5-year geopolitical-forecasting exercises. Here we extend this competition to 10-to-25 years. Although definitions of long-range tend to be conveniently elastic, Skeptics expect such exercises to become quixotic quickly whereas Meliorists are wary of broad generalizations. They expect pockets of predictability and unpredictability that create numerous chances for each school to claim vindication. In the realm of world politics, there are plenty of reasons for doubting the feasibility of long-range forecasting (Gaddis, 1992, 2019; Jervis, 2009) and plenty of anecdotes about spectacular forecasting flops: wars, revolutions, depressions, and technological trends that "experts" got wrong.

Given how pessimistic even optimists are in this debate, the current study sets its sights on a modest, proof-of-concept goal: finding any topics on which long-range geopolitical forecasters have a sporting chance to do well enough to persuade Skeptics to rein in extreme indeterminacy claims and work with Meliorists on documenting boundary conditions. We do not expect to settle the big debates but do hope to improve these debates. Discovering systematic (non-anecdotal) evidence of long-range accuracy should move the conversation away from either-or dichotomies toward a case-by-case weighing of claims and counter-claims.

To give forecasters this sporting chance, we focused on cumulative-risk judgments of slow-motion variables with low base rates of change and known antecedents: stability vs. change in national borders and nuclear-power status. To score forecasting accuracy, we drew on an old dataset from Tetlock's (2005) Expert Political Judgment project (henceforth EPJ) that consists of questions originally posed in 1988 and 1997 and only became fully resolvable in 2022.

EPJ's main focus was on shorter-range geopolitical forecasts that specialists in the late 20th century made of events inside or outside their domains of expertise—and that could be resolved in the next 1, 3 or 5 years. The findings were complex and reviewers eager to call out pseudo-expertise often exaggerated how poorly experts performed (Stevens, 2012; Gove, 2016). For instance, EPJ did find that experts toiling inside their domains struggled to out-perform trespassers from other fields—and had trouble beating base-rate extrapolation for leader turnover, policy changes and macro-economic indicators. But EPJ did not show experts' forecasts were no better than what we could have gotten from dart-tossing chimps. EPJ also found the accuracy of both specialists and generalists fell as a function of how distant events were in time—with the rate of decline depending on region, moment in history and outcome variable. EPJ did not however show the futility of forecasting beyond 3 or 5 years. It presented no data bearing directly on that controversy.

Although most EPJ forecasts were resolvable by 2003, some small-scale supplementary exercises tested longer-range forecasting skill: 10 and 25 years (resolvable in 2022) and even 50 and 100 years (resolvable in 2097). The ideal long-range question for a fair test of expertise falls in the Goldilocks zone of difficulty. It is far-fetched to suppose someone could predict, say, the President of the United States 25 years out. That sets up Radical Skeptics for easy wins. But it is less far-fetched to suppose someone could predict trajectories of slow-motion variables with low base rates of change: nation-state borders, nuclear proliferation or rankings of nation-states on GDP, life span or corruption. The challenge here becomes avoiding the flipside error of making it too easy to demonstrate skill. One did not need to be a geopolitical genius to see that Pakistan and North Korea posed greater nuclear proliferation risks than Ireland or Iceland.

Cumulative-risk forecasts of slow-motion variables are good Goldilocks-zone candidates. Structural engineers cannot predict exactly when houses will slide down hills but can offer educated guesses about how fast risk rises as a function of steepness of hill, depth of foundation, type of soil, and severity of storms. They can also measure variation in experts' accuracy over time and create wisdom-of-elite-crowd aggregates that, by tamping down noise, out-perform the individuals from whom the aggregates were derived (Kahneman et al., 2021). To be sure, social science work on slow-motion variables in world politics is not as rooted in rigorous science as structural engineering. But it offers grounds for optimism that experts could do something analogous: (a) identify chronic risk-amplifiers and attenuators for nuclear proliferation (Bass, 2016; Bell, 2016; Carless et al., 2021; Cimbala, 2017; Kaplow, 2016; Solingen, 2007, 2018), secession (Levinson, 2017; Siroky & Abbasov, 2021) and interstate border disputes (Carter & Poast, 2017; Huth, 2009); (b) synthesize evidence into causal-propensity assessments for the initial five years and convert those assessments into probability judgments on a bounded zero-1.0 scale; (c) adjust their initial probability judgments as time widens to 10 and 25 years; (d) average individual forecasts into composites to further enhance accuracy.

Moderate Meliorists do however temper their optimism. They concede the potential for exogenous shocks—earthquakes along undiscovered fault lines—to throw off even top forecasters. And they recognize that expert communities are fallible and prone to groupthink, fads and self-promotion.

With these caveats, we propose three sets of Meliorist hypotheses: (a) correspondence hypotheses bearing on the accuracy of experts and sophisticated generalists in short and long-range forecasting exercises; (b) coherence hypotheses bearing on the connections between empirical accuracy of forecasts and the logical rigor with which forecasters make cumulative-risk judgments; (c) viewpoint-variable hypotheses bearing on the connections among ideology, forecasts, and forecasting accuracy. Each set of hypotheses is purely correlational—and rests on the working assumption that the present contains clues to the future that observers with the requisite skills and mindsets are better at spotting. And for each set, Radical Skeptics offer an emphatic counter-hypothesis: the null hypothesis.

But before turning to hypotheses though, we should be explicit about the limitations of the current study, which was never a priority in the larger EPJ project. Few expected to find forecasting skill 25 years out, so the posing of such questions was sporadic, opportunistic and not woven into the overall design. As a result, the study has many methodological shortcomings: small sample sizes, inadequate measures of expertise, a flawed probability scale, and a rushed schedule that gave forecasters little time to deliberate. In this light, it is surprising we found anything worth writing up three decades later. The main value of the study is for jumpstarting conversations about follow-up work, not for delivering knock-out blows to schools of thought.

Correspondence Hypotheses Linked to Individual Differences in Expertise. The spotlight is on how closely subjective-probability forecasts of experts and non-experts correspond with objective reality increasingly far out in time (5, 10, and 25 years) in two domains: nuclear proliferation and border change/secession. Ideally, we could test expertise hypotheses by using a knowledge-and-skills litmus test of professional competence, analogous to board-certification exams for doctors and lawyers. We could then pinpoint who possesses relevant nomothetic knowledge—durable principles of causality about, say, centrifuges and highly enriched uranium—or case-specific knowledge—like satellite photos of North Korean reactors. But we lacked the means to do this and had to rely on a cruder operational definition that treated forecasters as experts if they had been educated at the post-graduate level in relevant disciplines and if they saw the topic as central to their professional identity.

Correspondence Hypothesis 1 expects experts to add predictive value—although how much hinges on whether experts can obtain the accuracy feedback critical for translating their professional knowledge into accurate judgments (Bolger & Wright, 1984; Keren, 1997). Even experts adept at putting current events in historical context will often be unable to translate those insights into accurate forecasts due to the opacity of the causal drivers in play.

Correspondence Hypothesis 2 expects forecasting accuracy, for experts and non-experts alike, to decline as a function of temporal distance. Events 25 years out will be harder to predict than events 10 or 5 or close-to-zero years out. But accuracy will fall off more slowly for experts who demonstrate a firmer command of predictively useful knowledge in the early periods.

Correspondence Hypothesis 3 stresses the dangers of déformation professionnelle. Credentialed experts may claim to be objective truth-seekers but they are corruptible. They have career incentives to exaggerate the value of guild-defining knowledge. Expect nuclear proliferation experts to see high risk of nuclear weaponry spreading and specialists in identity politics to see high risk of border conflicts. Expertise will add no predictive value if for every True-Positive (hit) forecast, they make many False-Positive (crying wolf) forecasts. Balancing these arguments, the third hypothesis predicts that experts will retain a predictive edge but it will be eroded by a tendency to make False-Positive judgments that highlight their own importance. Prudent experts realize crying wolf too often undermines their collective credibility.

Correspondence Hypothesis 4 focuses on robustness checks: How much does the accuracy scoreboard change when we recode the outcomes used to score accuracy in response to complaints of varying historical plausibility? We explored two types of adjustments. Close-call counterfactual adjustments are concessions to the indeterminacy of life, a recognition that things could have worked out otherwise. History is a stochastic process that yields probability distributions of outcomes (Tetlock & Belkin, 1996; Tetlock, Lebow & Parker, 2006). Controversy adjustments are admissions that researchers failed the clairvoyance test and did not specify resolution criteria clearly enough that a clairvoyant, handed the question, could answer without asking us to clarify, say, it means to be a "formally declared nuclear power." The fourth hypothesis posits: (a) expertise will cease to predict accuracy, or even flip in sign, when we replace actual outcomes with counterfactual ones that elite opinion of the day discounted but non-specialists' took seriously (e.g., nuclear war between 1997-2022); (b) expertise will gain predictive power when we replace actual outcomes with counterfactual ones that elites saw as more worrisome than did generalists (e.g., Iran a nuclear power before 2022).

Correspondence Hypothesis 5 focuses on aggregation algorithms for distilling the "wisdom of the crowd," a topic with a long history in the forecasting literature (Surowiecki, 2005). If forecasters work with roughly the same information but make random mistakes (noise) in converting that information into predictions, it should be possible to improve accuracy with measures of central tendency that cancel out errors of over- and under-estimation (Satopää et al., 2022a). But if forecasters base their predictions on different sets of information, we need more complex aggregation techniques to harness the insights dispersed among individuals. Merging dispersed information requires knowledge of forecasters' prior belief (Dietrich, 2010; Satopää, 2022b). In practice, the prior belief is often given by the base rate. The forecasters then form their predictions of a given event by updating the prior belief with distinctive information they possess. By analyzing how much forecasters update, sophisticated aggregators, like the Regularized Bayesian Aggregator (Satopää, 2022b), can begin to capture how information is dispersed among forecasters and combine the predictions into a more powerful composite. It follows that the accuracy ranking of aggregation tools will depend on why forecasters disagree: If largely due to noise, simple averaging will suffice; if due to information asymmetry, tools like the Regularized Bayesian Aggregator will have an edge.

Coherence Hypotheses: Individual Differences in Cumulative Risk Judgments. The focus shifts from getting it right to thinking the right way about the logic of cumulative risk. One source of judgmental error is noise: random variation in how people translate vague hunches onto probability scales (Kahneman et al., 2021). Another source of error is systematic bias: in this case, the well-replicated tendency to under-estimate how fast risk adds up over time (Fischhoff et al., 1993; Tversky & Kahneman, 1983). Researchers are free to ask for probability judgments but there is no guarantee people will oblige. They may do something that feels more natural: make causal-propensity attributions grounded in hunches about capabilities and intentions (Tetlock, 1998).

Causal-propensity judgments are tricky to translate into probabilities for two reasons. First, propensities are numerically unbounded whereas probabilities are confined to the zero-to-one interval. This mismatch creates the propensity-to-probability conversion problem: how to map magnitudes of causal-propensities into probabilities. Second, propensities are static. Once people label an individual or regime aggressive, that attribution will stick (Nisbett & Ross, 1980). Absent new evidence, many tacitly assume the entity will remain aggressive. Whatever propensity-probability conversion people perform for the first period will anchor future judgments, producing temporal-scope insensitivity.

Logical-Coherence Hypothesis 1 predicts that more probability-savvy forecasters will estimate the cumulative risk of event *X* over *T* periods by using an informal version of the formal compounding formula familiar to actuaries:

$$P(X) = 1 - (1 - f_t)^T$$

where f_t is the forecasted probability of an event in a single time period. If forecasters see a 1% chance of nuclear war each year and consider the outcomes each year as independent, the chances of at least one nuclear war in the next 100 years are 63.4%. Note we are not assuming 1% is correct, just that for a stationary time series, the correct way to combine 1% estimates yields 63.4%.

By contrast, less probability-savvy forecasters will resort to shortcuts when sizing up risk. Some heuristics, like the additive one, will cause over-estimation.[1] Forecasters simply sum the 1% probabilities across 100 years and put a 100% likelihood on nuclear war in the next century:

$$\sum_{t=1}^{T} f_t$$

Other heuristics, like averaging, cause under-estimation of cumulative risk.[2] Forecasters average the 1% probabilities each year—and conclude the risk for the entire century must also be 1%:

$$\frac{1}{T}\sum_{t=1}^{T} f_t$$

Logical-Coherence Hypothesis 2 links the rigor of forecasters' judgment process to the accuracy of their forecasts. Assuming forecasters attach well-calibrated probabilities to the initial five-year period and a stationary time series of true outcome probabilities across the 25- year span, forecasters who base their later forecasts on the compounding formula will out-perform those using bounded-rationality heuristics. But there is potential for offsetting errors. Forecasters who start with too high probabilities in the initial period may be spared the accuracy consequences of that mistakes if they use an anchoring heuristic that makes later judgments insensitive to temporal scope (sluggish judgments across 5/10/25 years: say, .5/.6/.6). And forecasters who start too low may be spared the consequences if they use a heuristic that is hypersensitive to temporal scope (judgments leap from, say, .2 to .4 to 1.0).

[1] If $f_t = f$ for all t, then each time period adds f. Given that $f(1-f)^{t-1} \leq f$, we have that $1-(1-f)^T \leq Tf$.

[2] If $f_t = f$ for all t, then the average forecast is f and $(1-f)^T \leq (1-f) \Leftrightarrow f \leq 1-(1-f)^T$.

Viewpoint Hypotheses: Individual Differences in Perspectives. Expertise is only one of a vast set of possible individual difference predictors of accuracy. Forecasters vary not only in their command of technical knowledge but in their less fact-constrained opinions about how the world works and in their tolerance for False-Positive errors. Ideally, researchers would have access to an array of measures of ideology, theoretical affinities, and cognitive styles as in the original EPJ project. Again, lacking the necessary resources, we could only include two opinion scales in our 1997 measurement net to capture individual differences in the degree to which:

(a) forecasters who saw nuclear deterrence as inherently fragile (e.g., the USA and USSR came close to nuclear war in Cold War) will see nuclear proliferation and war as likelier than their colleagues who saw nuclear deterrence as robust. Which theoretical camp proves more accurate hinges on whether "fragility" is the true state of the world and nation-states react to the true state by accelerating their own nuclear programs or by supporting anti-proliferation and disarmament initiatives;

(b) forecasters who viewed world politics as undergoing a paradigm shift in which powerful macro forces are overwhelming the old Realpolitik ground rules will see nuclear weapons as less useful than colleagues who expect no such transformation. Whether the paradigm-shifters prove more accurate than traditionalists hinges on whether history has reached an "end-of-history" turning point (Fukuyama, 1989, 1995) or the Realpolitik ground rules of recorded history can hold for another quarter century.

Note we are expecting viewpoint variables to predict forecasts, not accuracy. Whether fragility or end-of-history theorists prove more accurate than their sparring partners hinges on whose theoretical diagnoses are closer to the true state of the world.

Purely Exploratory. In 1997 only, we elicited ultra-long-range forecasts on nuclear war that stretched out beyond 25 years to 50 and 100 years, which means we must wait until 2097 for final resolution. We can still though assess accuracy in the first 25 years as well as correlates of accuracy–and make plausible extrapolations.

Method

Context. The study drew on EPJ studies of experts' probabilistic forecasts, elicited in 1988 or 1997, on the trajectories of slow-motion variables over the next 5, 10 and 25 years. In 2022, it became possible to assess forecasting accuracy on two such variables: changes in nuclear-power status of nation-states and changes in nation-state boundaries. To qualify as slow motion, a variable had to have low base rates of change over long periods (less than 25% change over 25 years) but the base rate could not fall too close to zero (less than 5%). We only used exercises that had at least 25 forecasters and at least 25 forecasts per-participant.

Forecasting Questions. The long-range forecasting questions in both 1988 and 1997 were:

- Looking ahead 5/10/25 years, how likely do you think it is that each of the following nations will officially declare itself in possession of nuclear weapons (a declaration backed by evidence of at least one verifiable test explosion)?

The long-range forecasting questions that were only asked in 1997 were:

- Looking ahead 5/10/25 years, how likely do you think it is that the borders of each of the following nation states will change due to:
  a) A secessionist movement that breaks off and successfully establishes control over territory previously under the control of the nation-state (successful secession requires recognition by at least one other nation-state).
  b) A change in interstate boundaries brought about by military force and/or political negotiations (boundary change that must be recognized by at least one other nation-state).

These two sets of questions are henceforth abbreviated as NP for nuclear proliferation and BCS for Border Change/Secession.

The ultra-long-range question was about nuclear war and only asked in 1997:

- Looking ahead 10/25/50/100 years, how likely do you think it is that one or more nuclear weapons will be used in combat (not just a warning shot—a targeting of military or civilian targets)?

Cumulative-Risk Probability Scaling. Forecasters used the same 11-point, zero-to-one, subjective probability scale as in other EPJ exercises, with equal 0.1 spacing between levels (0, .1, .2, … , .9, 1). The other EPJ exercises were not however focused on events with base-rates as low as here (indeed, nuclear war has a zero base rate). Some participants complained about the insensitivity of the scale to low probabilities in the zero to 0.1 range. In response, we gave this guidance: "if you feel the probability is closer to zero (virtually impossible) than to 0.1, please round down to zero. We understand that distinctions this fine-grained are hard to make and, given the limited time, ask you to make these judgment calls quickly as well as carefully." With benefit of hindsight, we should have adopted a method, like Karger et al.'s (2022) nonparametric continuous-probability elicitation-technique (see Discussion).

Forecasters used the 11-point scale to make cumulative judgments across 5, 10 and 25 years and were given this guidance: "Please remember we are asking for cumulative-risk judgments. So if you think the likelihood of something happening is 10% in the next five years, it is logically necessary that the likelihood of that thing happening in the next 10 years must be at least 10%. Cumulative risk estimates can never go down over time. But we need your judgment call about how slowly or rapidly cumulative risk rises. You may see risk remaining very low throughout the 25-year period—or as rising steeply. We need your rapid-fire, best educated guesses."

Probability Scoring Rules. We relied on two widely used *proper scoring rules* for gauging goodness of fit between subjective probabilities and objective reality: quadratic (a.k.a., Brier)—our primary scoring rule—and logarithmic, which we use to supplement the implied conclusions from the Brier score. A rule is "proper" if it incentivizes forecasters to give their true best estimate. This discourages the quite common real-world practice of fact-value conflation and of shading one's forecasts to avoid the more distasteful mistake (Tetlock, 2005). Indeed, using improper scoring rules may result in grossly misguided inferences about forecasters' skills (Gneiting and Raftery, 2007).

Brier Scoring. The quadratic Brier equation is: $(f-x)^2$, where x (a dummy variable) equals 1 if the forecasted outcome occurs or 0 otherwise, and f is the forecaster's probability of occurrence. Forecasters receive the ideal score when they always assign predictions of 0 to outcomes that do not occur (so, $(f-x)^2=(0-0)^2=0$) and always assign predictions of 1 to outcomes that do occur (so, $(f-x)^2=(1-1)^2=0$). For binary events, Brier scores range from 0 (perfect score: no gap between probability judgments and reality) to 1.0 (worst possible score: maximum probability-reality gap). For instance, if a forecaster put a 90% chance on an event that occurred, the Brier score would be $(.9 - 1)^2 = .01$. If the event did not occur, the Brier score would be $(.9 - 0)^2 = .81$. As a benchmark, a recurring forecast *f* of .5 would always get a Brier of .25.

We use two decompositions of Brier. The first breaks Brier into calibration (CAL) and refinement (REF):

$$BS = CAL + REF = \frac{1}{11}\sum_{k=1}^{11} n_k(f_k-\bar{o}_k)^2 + \frac{1}{11}\sum_{k=1}^{11} n_k\big(\bar{o}_k(1-\bar{o}_k)\big)$$

where *k* indexes the eleven possible prediction values in our discrete probability scale $f_k \in \{0, 0.1, \ldots, 0.9, 1\}$, n_k denotes the number of times the forecaster predicted f_k, and $\bar{o}_k$ denotes the observed relative frequency of the events for which the forecaster predicted f_k. *Calibration* (also known as *reliability*) describes how well the forecaster's probabilities align with empirical frequencies of the events and hence can be interpreted as probabilities (in the frequentist sense). Here calibration is measured with the weighted average of the mean-square differences between proportion of occurrences in each probability category and probability value of that category. A score of zero is perfect calibration. *Refinement* gauges skill at sorting outcomes into the occurrence and non-occurrence categories, and is related to the Area Under the Curve (AUC) metric from signal detection theory. As with calibration, lower refinement scores contribute to better (lower) Brier scores.

Calibration and refinement tend to be correlated but one can get a good calibration but poor refinement score if one assigns only the base-rate probability to every forecast. In this case, the observed event frequency matches the forecast yet there is no skill in separating occurrences from non-occurrences. Likewise, one can also get good refinement and poor calibration scores by being consistently and dramatically wrong. High-calibration-and-refinement forecasters give us the best of both worlds: a realistic sense of how well they can differentiate lower from higher probability events.

The second decomposition includes the calibration (CAL), uncertainty (UNC), and resolution (RES) components of Brier (Murphy, 1973; Yaniv et al., 1991):

$$BS = CAL + UNC - RES = \frac{1}{11}\sum_{k=1}^{11} n_k(f_k-\bar{o}_k)^2 + \bar{o}(1-\bar{o}) - \frac{1}{11}\sum_{k=1}^{11} n_k(\bar{o}_k-\bar{o})^2$$

where $\bar{o}$ is the overall occurrence frequency or base rate. The uncertainty term, often called, entropy, does not depend on the predictions. It is zero if the event always occurs or never occurs, and reaches a maximum when $\bar{o} = .5$. The resolution term captures skill at beating the base rate.

Logarithmic Scoring. Forecasters get a score of *-ln(f)* if the event occurs—or *-ln(1-f)* if it does not occur. Suppose two forecasters assign probabilities of .8 and .3 to a non-occurrence. Their respective log scores are –ln(.2) = 1.61 and –ln(.7) = .36. Forecasters who put a zero probability on an event that occurs—or 1.0 on an event that does not occur—get a score of negative infinity, a loss of all credibility for eternity. In contrast, the Brier score is always bounded between 0 and 1. Log scoring is therefore tougher on extreme over-confidence than quadratic scoring, yet log scoring and Brier scoring differ far less in how they treat less dramatic mistakes. To eliminate the negative infinity problem and be consistent with the rounding-down-or-up instructions given forecasters, we reset forecasts of 0 to .01 and forecasts of 1 to .99 for log scoring.

Retrospective Counterfactual Questions. In 2022 and in consultation with colleagues, we made informal judgments on junctures at which history could have gone down different paths—an exercise to estimate the robustness of conclusions across alternative runs of history. The counterfactual-plausibility scale ran from .001 (Extremely Implausible) to .999 (Extremely Plausible), with midpoint of .50 anchored as Moderately Plausible—and asks respondents to "look back over the last 25 years, from 1997 to 2022 and identify any what-if scenarios that they saw as falling in the moderate plausibility range (.25-.75) or higher. We asked about the plausibility of supposing that nation-states that became declared nuclear powers not doing so (Pakistan, India, North Korea) and of supposing that nation-states that did not become nuclear weapons powers doing so (Iran, Iraq, Libya, Syria, Saudi Arabia, Japan, South Korea, Taiwan, Germany). Only one moderately plausible counterfactual candidate emerged: Iran.

Forecaster Samples and Tasks. Table 1 summarizes the types of data collected in 1988 and 1997, including sample sizes of expertise, viewpoint variables, and forecasting questions. The 1988 study classified participants as subject-matter experts on nuclear proliferation based on responses to a true-false dichotomous item: "A good portion of my professional career has involved the study of either nuclear weapon systems, nuclear strategy or issues of nuclear proliferation." Four of thirteen participants agreed and were classified as proliferation experts in 1988. In 1997, we posed the same question but used a 5-point scale, anchored by "not at all true" (1) and "very true" (5), with the request to check 4 or 5 "only if you have written articles or other analytic products on the topic." Eleven of 34 forecasters checked 4 or 5 for the NP expertise question in 1997.

We identified 1997 participants as experts on border-change/secessionism if they checked 4 or 5 on the same 5-point scale, now anchored by the request to check "only if you have written articles or other analytic products on secessionism, civil war and interstate boundary disputes." Using this definition, we identified 9 boundary experts out of 34 forecasters in 1997. Given the different measures in 1988 and 1997, we simplify analyses that draw on both datasets by using the (admittedly inferior) dichotomous measure. We note that NP and BCS expertise were not mutually exclusive, and forecasters could be classified as both. We analyze each domain through the lens on the singular, relevant expertise.

As Table 1 notes, 1988 data collection only involved nuclear proliferation forecasts. Border-change/secession questions were only asked in 1997—and expertise on that topic was only assessed in 1997. For the 1997 sample only, we also collected responses to four opinion items:

- Item 1: Looking back on the history of the Cold War, how much do you agree or disagree with the claim that the USA and USSR came very close to a nuclear war? Scale values: 1 = strongly disagree; 9 = strongly agree; 5 = unsure.
- Item 2: How often do you think the US and USSR came very close to nuclear war? Scale values: 0, 1, 2, or "3 or more times".
- Item 3: Some argue that nuclear deterrence is inherently fragile (easily shattered by human irrationality and chance events—so not a reliable safeguard against nuclear war). Others argue that nuclear deterrence can be robust with clear communications, tight command and control, and mutual assured destruction. Which comes closer to your views: the first or second position? Scale values: 1 = extremely fragile; 9 = extremely robust; 5 = unsure.
- Item 4: Some argue we live in a transformational period in world politics and the old realist maxim ("the strong do what they will and the weak accept what they must") is becoming irrelevant. Others argue that world politics will continue to be dominated by a power-calculus logic. Which comes closer to your views: the first or second position? Scale values: 1 = lean strongly toward first position; 9 = lean strongly toward second position; 5 = unsure.

To respect forecasters' time, we stressed we were only asking for quick-assessment, best guesses for each nation-state for each 5/10/25-year period. Most respondents completed the proliferation judgments in about one hour and took slightly longer for the border-change/secession questions. To allay concerns about professional embarrassment, we promised anonymity. The instructions also stressed that the task was to judge cumulative likelihood over time—so it would be logically impossible for an event to be less likely to occur in the next 10 years than it is in the next 5 years. Cumulative likelihoods cannot fall over time. Finally, data was also collected in 1997 that asked about the probability of nuclear war in the next 10, 25, 50, and 100 years.

Selecting Entities for Forecasting. Assessing base rates of change is tricky due to ambiguity about both the numerator (how often nuclear-power status or boundary control changes) and the denominator (how to define the reference class of nations). It is easy to shrink the base rate of change by including all 200+ nation-states in our sample. Iceland and Fiji are extremely improbable nuclear powers—and Norway and Sweden are extremely unlikely to go to war. So defining the candidate space was a key design decision. For the proliferation candidate set, we based our judgment calls about long-run proliferation risks on: (a) evidence of an ongoing weapons development program (e.g., North Korea, Pakistan, Iran, Iraq, Libya); (b) evidence of a past program now abandoned (e.g., South Africa, Brazil, Argentina, Sweden, Ukraine, Kazakhstan); (c) the ease of making a prima facie case that nation-states had the capacity to develop a nuclear weapon. India and Israel were complicated cases. Both had already detonated an atomic bomb—India in 1974 and Israel in 1967—but neither had officially declared itself a nuclear-weapon-possessing state. The India question was therefore: how likely is India in the next 5/10/25 years to transition from its "smiling Buddha" policy (of not weaponizing an existing capability) to an explicit deterrence posture of weaponizing warheads and conducting demonstration tests? (India conducted five tests in 1998—which was followed by six Pakistani nuclear tests.) The Israel question was: how likely is Israel in the next 5/10/25 years to abandon its thinly veiled policy of strategic ambiguity (not admitting its nuclear arsenal) to formally declaring itself a nuclear power?

In the sample of proliferation suspects there were 21 entities judged in 1988 and 21 entities judged in 1997. The total sample included: India (only in the 1988 elicitation—a policy shift question), Israel (only in 1988—a policy shift question), Pakistan, North Korea, Iran, Iraq, Saudi Arabia, Germany, Japan, South Korea, Taiwan, Vietnam, Egypt, Turkey, Libya, Ukraine (only in 1997), Kazakhstan (only in 1997), South Africa, Brazil, Argentina, Sweden, Nigeria, and any sub-national actor/terrorist group. There were 3 changes in nuclear status from 1988 through 2022: in 1998, when India and then Pakistan conducted multiple nuclear tests, and in 2006 when North Korea conducted its first test. India was only asked about in the 1988 elicitation, but Pakistan and North Korea were questions in both 1988 and 1997. Therefore, India counted as positive outcome for the 1988 10-year and 25-forecasts, Pakistan counted as a positive outcome for the 1988 and 1997 10-year and 25-year forecasts and for the 5-year 1997 forecast, and North Korea counted as a positive for the 1988 25-year forecast and the 1997 10- and 25-year forecasts. So, for example, the base rate of change for the 25-year forecast has five positive outcomes (India once and North Korea and Pakistan both twice) over 42 total questions: 5/42 = 12% over 25 years for the entities that were deemed in 1988 and/or 1997 to be plausible long-run proliferation risks.

For inclusion in the border-change set, judgment calls rested on whether attentive readers of the news could currently see or plausibly imagine: (a) serious territorial disputes arising with neighboring states (Russia-Ukraine; Ethiopia-Eritrea/Somalia, China-Taiwan, …); (b) secessionist movements arising within states (Sudan, Canada, Iraq, …). Again, we were aiming for questions in the Goldilocks zone: challenging but not impossible.

For the 40 national-boundary questions, we defined change to encompass either de jure change (internationally recognized) or de facto change (compelled by military force but not widely recognized). The base rate of change over 25 years was 9/40 = 23%, with verified change resolutions for questions about the secession for Sudan and Ethiopia and verified border changes for Ethiopia-Eritrea, Ethiopia-Somalia, Armenia-Azerbaijan, Russia-Georgia, Russia-Ukraine, Serbia-Croatia, and Serbia-Kosovo. The possible boundary disputes included: Iraq-Iran, Armenia-Azerbaijan, Ethiopia-Eritrea, Ethiopia-Somalia, Ukraine-Russia, Kazakhstan-Russia, Estonia-Russia, Latvia-Russia, Lithuania-Russia, Belarus-Russia, China-Russia, Japan-Russia, Serbia/Croatia/Bosnia/Kosovo, China-India, China-Vietnam, and China-Taiwan. Secessionist possibilities included: Russia, Kazakhstan, Indonesia, India, Pakistan, Canada, Spain, United Kingdom, Nigeria, Iraq, Iran, Syria, Lebanon, Sudan, Afghanistan, Ethiopia, and Sudan.

Results

Preliminaries to Data Analysis. Beyond the small sample sizes, the data had three other less-than-ideal properties. First, the 11-point probability scale forced forecasters to round their judgments between zero and .1, either down to zero or up to .1. Given they were judging low-frequency events, the result was that 65% of all judgments were zero. This distortion has little effect on Brier scores but a big one on log scores whenever zero-assigned events occur (as noted previously, we reset all 0 values to .01 and all 1.0 values to .99 for log scoring). Second, there was heterogeneity of variance across nations and time, a problem for *t* tests. To compensate, we used the Satterthwaite-Welch correction which contains Type 1 error by conservative adjustments to degrees of freedoms. Third, forecasters did not answer about 7% of the questions—so we used list-wise deletion to reduce data loss.

Correspondence Hypotheses 1 and 2: **Expertise and Temporal Effects on Accuracy.** In addition to the naïve Brier score threshold of .25 obtainable by forecasting 50% for each event, we used two more realistic but simple baselines for gauging human performance: predict-no-change (every forecast is 0) and base-rate extrapolation which always predicted the 5/10/25 year event frequencies for each period. These base-rate forecasts have the unfair advantage of being informed by perfect knowledge of how often things happened in the next 25 years—knowledge no observer could have had in 1988 or 1997, so beating these forecasts is a nontrivial achievement.

Table 2 shows event base rates and accuracy scores (both Brier and log scores) of the nuclear proliferation forecasting algorithms, experts and non-experts. Proliferation experts beat both no-change and base-rate-extrapolation algorithms across the latter two time periods by wide margins. Non-experts mostly cleared these algorithmic benchmarks but by much narrower margins. To explore skill differentials in Brier scoring, we averaged experts' and non-experts' Brier scores for each nation-state and period (n = 126) and computed paired t-tests and Cohen's d effect sizes. Experts beat Brier scores based on time-relevant base-rates (t(125) = 2.44, p = .02, d = 0.22) and the averaged Brier scores of non-experts (t(125) = 4.50, p < .001, d = 0.40). To test time-horizon and expertise effects, we regressed forecasters' averaged-across-questions Brier scores on time and proliferation expertise. The intercept coefficient (non-expert error at timeframe 0) was significant (B = 0.03, SE = 0.01, p = .02), as were timeframe (B = 0.002, SE = 0.001, p = .004) and expertise (B = -0.02, SE = 0.01, p = .04) with R^2 = 0.05, F(2, 249) = 6.36, p = .002. In short, long-run proliferation questions were harder but expertise conferred an advantage.

To isolate the source of experts' edge in the proliferation domain, we decomposed Brier scores into calibration and refinement. Across time periods, experts did better on both components of the Brier, with their biggest advantage on refinement (Avg. BS for experts = .034; Calibration = .002; Refinement = .032 vs. Avg. BS for non-experts = .057; Calibration = .003; Refinement = .054). The three-part decomposition yields a similar pattern: Experts delivered better resolution (Avg. BS for experts = 0.034; Calibration = 0.002; Uncertainty: 0.078; Resolution: 0.046 vs. Avg. BS for non-experts = 0.057; Calibration = 0.003; Uncertainty: 0.076; Resolution: 0.023).

Table 3 shows assigned probabilities to occurrences and non-occurrences in the proliferation domain, with forecast counts and standard deviations in parentheses. Experts consistently bested non-experts at assigning higher probabilities to occurrences than to non-occurrences.

Table 4 shows that unlike NP experts, experts in the BCS domain were not better forecasters than non-experts. Again we averaged Brier scores for each group across 120 questions (nation-state(s) and time period) and computed a paired t-test and Cohen's d. Although both non-experts and experts beat base-rate extrapolation (t(119) = 2.40, p = .02, d = 0.22; t(119) = 2.75 p = .007, d = 0.25, respectively), the difference between experts and non-experts was not significant (t(119) = 0.82; p = .41, d = 0.08). Applying the same regression methods to test the expertise and thinking-in-time hypotheses for BCS outcomes, neither the intercept coefficient (non-expert error at timeframe 0; B = 0.03, SE = 0.02, p = .22) nor the expertise coefficient emerged as significant (B

= 0.005, *SE* = 0.02, *p* = .85) but timeframe did (*B* = 0.005, *SE* = 0.001, *p* = .001), with R^2 = 0.05, *F*(2, 237) = 5.60, p = .004.

We compared two-part decomposed Brier scores across all periods for BCS experts (Avg. BS = .097; Calibration = .002; Refinement: .094) and non-experts (Avg. BS = .092; Calibration = .002; Refinement: .09)—and found nearly identical calibration and refinement. The three-part decomposition yields the same result (Avg. BS for experts = .097; Calibration = .002; Uncertainty: .128; Resolution: .033 vs. Avg. BS for non-experts = .092; Calibration = .002; Uncertainty: .128; Resolution: .038). Finally, Table 5 shows average probabilities assigned to occurrences versus non-occurrences. Overall, BCS questions were harder than proliferation questions, with Brier score error growing 2.5 times faster than in the proliferation domain (regression coefficients of 0.005 vs. 0.002) and BCS expertise did not yield any accuracy advantage. Algorithms and humans alike performed worse in the BCS domain.

Correspondence Hypothesis 3: False-Positive Tolerance. As Brier decompositions underscore, there are many ways to get a bad forecasting score. For instance, forecasts may be, on average, too far above or below the base rates. Table 6 and 7 present the average differences between each forecaster's average prediction and the base rates in the proliferation domain and border change/secession domains. Negative values indicate under-estimation; positive values, over-estimation. Bracketed values give the 95% bootstrap confidence intervals.

Forecasters over-estimated risks of nuclear proliferation but, contrary to Hypothesis 3, experts did so less than non-experts. Hypothesis 3 also fails to capture the data in the BCS domain. Here forecasts were, on average, too low in all time frames. The underestimation is significant for 5-year but not 10- or 25-year predictions—and experts and non-experts differ only slightly.

To further analyze the balancing of errors, we calculated simple binary classification and error rates by focusing on probabilities less than or greater than 50% and marking each forecast as a correct classification if: (a) the forecast was below 50% and the outcome did not occur; or (b) the forecast was above 50% and the outcome did occur. The denominator for the correct classification rate was the total number of qualifying forecasts. Likewise, false positives were considered where the forecast was above 50% and the outcome was zero, and false negatives were considered where the forecast was below 50% and the outcome was one. The false positive rate was calculated using the number of forecasts that exceeded 50% as the denominator whereas the false negative rate was calculated using the number for forecasts that were less than 50% as the denominator.

In the NP domain (across time), experts had a correct classification rate of 96% (825/858), a false positive rate of 23% (15/66), and a false negative rate of 2% (18/792). Non-experts had a correct classification rate of 94% (1,692/1794), a false positive rate of 45% (37/83), and a false negative rate of 4% (65/1711). In the proliferation domain, experts did better. Their FP/FN rates were half those of non-experts. These experts were better at spotting risks, not just at crying wolf.

In the BCS domain, experts had a correct classification rate of 90% (1026/1140), a false positive rate of 24% (18/74), and a false negative rate of 9% (96/1066). Non-experts had a correct classification rate of 89% (2134/2389), a false positive rate of 27% (41/151), and a false negative rate of 10% (214/2238). In the BCS domain, classification rates were lower and false negatives were notably higher than in the proliferation domain.

Correspondence Hypothesis 4: Robustness Checks. Controversy adjustments respond to claims we misread reality and miscoded events as non-events (e.g., disputes over when a de facto but undeclared nuclear power declared itself). Close-call-counterfactual adjustments respond to claims that history nearly yielded alternative outcomes (e.g., disputes over how much accuracy scores pivot on close calls—on whether we recode a secession that almost happened as an actual event).

The major controversy adjustment in the proliferation domain assessed the impact of treating India as a declared nuclear power in 1988-93 rather than waiting for the multiple explosions of 1998. The adjustment did not change experts' and non-experts 5-year Brier scores (still .02 and .03, respectively). The major close-call counterfactual adjustment recoded Iran as a nuclear-weapon state between 2008-2022, which is shown in Figure 1. This produced a slight uptick in the benefits of expertise (from $d = 0.40$ in real world to $d = 0.43$ in counterfactual world). In addition, we tested the impact of a more extensive rewrite of history (also in Figure 1): a counterfactual nuclear arms race in the Middle East with weapons spreading beyond Iran to Saudi Arabia, Iraq and Libya. While Brier scores increase dramatically, this does reduce experts' edge over non-experts (Brier differential: .03 vs. .02), but the edge remains ($t(125) = 4.50$, $p < .001$, $d = 0.40$ in real world; $t(125) = 4.02$, $p < .001$, $d = 0.36$ in counterfactual world).

Shifting to border-change/secession, we tested two controversy adjustments: the impact of flipping the outcome of the Ethiopia-Eritrea conflict (from border change to no change over all three time periods) and the Sino-Indian border outcome (from no border change to change for the 25-year time horizon). The first adjustment had almost no effect on Brier scores; the second one degraded Brier scores roughly equally for experts and non-experts from .15 to .17 for 25-year forecasts. The close-call counterfactual adjustments explored a rewrite of history in which three failed secession efforts singularly succeeded (shown in Figure 2): Quebec, Scotland and Kurdistan. Brier scores increased slightly, yet the expert/non-expert gap remained non-significant.

Correspondence Hypothesis 5: Aggregation. We expected the accuracy ranking of aggregation techniques to depend on why forecasters disagree: if largely due to noise, the average or other measures of central tendency will suffice but if largely due to information asymmetry, more complex techniques like the Regularized Bayesian Aggregator will have an edge.

Given that aggregate predictions often outperform individual forecasters, Figure 3 starts the analysis by considering crowds of different sizes and comparing the Brier scores[3] of the base-rate extrapolation, simple averaging (AVE), and Regularized Bayesian Aggregator (RBA)[4] with prior beliefs given by the time frame and domain specific base rates. For each period and domain, we sampled 500 crowds of a given size, aggregated each crowd's predictions, and plotted the overall average score. Note that, in 1988, as few as three experts made forecasts on certain nuclear

[3] Results based on the log score are qualitatively similar and deferred to the Supplementary Material.

[4] We use the implementation in the R-package braggR that is freely available on CRAN: https://CRAN.R-project.org/package=braggR.

proliferation questions, so we terminated the aggregation at that point in the proliferation domain. In each sub-figure, the right-most points (above the label “All”) represent scores from aggregating all non-experts’ or experts’ predictions in our dataset.

The results show that in each domain and time frame, aggregating sufficiently many predictions outperforms the average-accuracy individual forecaster (left most point in each sub-figure) and base-rate extrapolation (dashed horizontal line). Aggregating more predictions typically improves accuracy but with diminishing returns. Previous research has found similar patterns (e.g., Ashton & Ashton, 1985; Palley & Satopää, 2023). Theoretically, this follows if the forecasts are modeled as independent and unbiased (e.g., Gaba et al., 2019). In our study, aggregation yields the largest benefits in the longest time frame of 25 years. Overall, Brier scores are lower in the nuclear proliferation domain, again confirming that predicting future nuclear proliferation was easier than predicting border change. With enough predictions in each scenario, RBA performs the best in 5- and 10-year prediction accuracy, whereas averaging has a slight edge in 25-year prediction accuracy. One explanation is that 5- or 10-year timeframes may be short enough to allow forecasters to bring specialized knowledge to bear, creating information asymmetry that can be leveraged by RBA. In the 25-year timeframe, however, the differences may stem largely from noise, which is eliminated efficiently by simple averaging (Satopää et al., 2022a). Finally, as before, expertise leads to better accuracy only in the nuclear proliferation domain. Indeed, in the border change/secession domain it is slightly better to aggregate non-experts’ than experts’ predictions, which highlights both the non-experts’ modest edge in the border change/secession domain.

Outperforming the average-accuracy individual is less impressive than out-performing all or most of the individuals included in the aggregation. Figure 4 shows the percentage of forecasters whose average Brier score is worse than that of the simple average and the Regularized Bayesian Aggregator. Both aggregators outperform most individuals in all time frames—regardless of whether we aggregate all forecasters’ predictions or only experts or non-experts. The fractions tend to increase as the time frame becomes longer and are higher in the border change than in the proliferation domain. In most contexts RBA outperforms a larger fraction of individuals than simple averaging does. This can be expected because RBA is designed to construct an aggregate forecast that is more informed than any individual in the crowd, whereas averaging and other measures of central tendency are pure noise-reduction mechanisms (Satopää et al., 2022a). Even though a few individuals outperform RBA in the 5- or 25-year frames, no one beat RBA if we aggregate either all forecasters’ predictions or only non-experts’ predictions made for either all time frames or the 10-year time frame.

Logical-Coherence Hypothesis in Proliferation and Border-Change/Secession Domains. Paired *t*-tests ruled out the crudest heuristic hypothesis that forecasters just assigned the same value over time. The mean difference between 5- and 10-year forecasts across domains was .04 ($t(2{,}171) = 22.92$, $p < .001$) and between 10- and 25-year forecasts was twice as large at .08 ($t(2{,}171) = 30.08$, $p < .001$). Forecasters passed this minimalist coherence test. They were somewhat scope sensitive.

Next, we tested how well the risk-compounding and summation-heuristic hypotheses captured forecasters’ judgments. Using actual 5-year forecasts as inputs, we computed: (a) the 10- and 25-

year forecasts that follow from applying the two formulas of compounding-risk and simple summation (capping sums at 1); (b) the average absolute differences (of 10- and 25-year pairs) between calculated and actual forecasts. If the triplet forecast across three periods was {.2, .5, .8}, the compound-risk triplet would be (.2, .36, .67), with a mean absolute difference (MAD) of 0.135 ($\frac{1}{2}(|.5 - .36| + |.8 - .67|)$), and the summation-heuristic triplet would be (.2, .4, 1.0), with a MAD of 0.15 ($\frac{1}{2}(|.5 - .4| + |.8 - .1|)$). A paired *t*-test showed the compounding model better fit the data ($t(2{,}171) = 18.08$, $p < .001$, $d = 0.39$) with a mean MAD of $M = .08$ ([.07, .08]) versus $M = .10$ ([.09, .10]) for the summation model.

Regression analyses of MAD values for the compounding model revealed that domain was significant ($p = .002$) but domain expertise was not ($p = .31$). The mean compounding model MAD value for proliferation was $M = .07$ ([.06, .08]) and for BCS was $M = .08$ ([.08, .09]). The accounting formula was thus a better fit than the summation formula, especially in the proliferation domain, though even here there was still notable discrepancy between the accounting-formula-assigned forecasts and actual forecasts for 10 and 25 years.

Figure 5 shows when forecasters strayed the most and least from the risk-compounding model. It displays average differences between model-derived and human forecasts for 10 and 25 years, split by domain expertise. Putting aside cases when the 5-year human forecast was 0 (and later model-derived forecasts had to be zero), most people making 5-year forecasts of 0.1 or higher fell below the diagonal—and proliferation experts' forecasts were closest to model-derived values. One interpretation is that these findings support two mutually reinforcing hypotheses: (a) people are relying on Kahneman's (2011) anchoring heuristic, finding it hard to escape the tug of their initial estimates and under-adjust; (b) as Hypothesis 2 posited, more bias resistant and logically coherent forecasters (who turned out to be proliferation experts) tended to be more empirically accurate. A competing interpretation is that the fault lies not with the forecasters but with the normative model which treats border-change/secession risk as stationary over time when it may actually rise at differential rates, sometimes slowing and other times accelerating. There may well be periods of history when this second interpretation is true but 1997-2022 is not one of them.

Viewpoint Variables and Nuclear War. We now turn to the viewpoint variables and nuclear war forecasts obtained only in 1997. As previous analyses showed, expertise predicted accuracy in the proliferation but not the BCS domain. These additional individual differences let us identify the viewpoints linked to proliferation forecasts as well as to ultra-long-term forecasts of nuclear war. Nuclear-proliferation has a low base rate but, since Hiroshima and Nagasaki, the base rate for nuclear war has been zero—which puts it in a category of its own. That was why we stretched "long-range" from 25 years to 50 and 100 years. But that made nuclear war problematic in another way: of the four periods only the 10 and 25 years were resolvable in 2022. The full story must wait until 2097. We can though tell part of the story.

Table 8 summarizes data on viewpoint variables when partitioned on the dichotomous definition of proliferation expertise. The first two variables show the averaged 5-scale responses of NP and BCS expertise, followed by the averaged centered forecasts (each forecast had the mean question forecast subtracted from it) and the averaged normalized Brier scores (each Brier score had the mean question Brier score subtracted from it, and then divided by the standard deviation of Brier scores). Proliferation experts saw less risk of nuclear war than their non-expert counterparts—and thus got better Brier scores.

Figure 6 shows the correlations among these variables, now using the 5-point scale responses for expertise. Shaded cells are significant at the .05 level. The four leftmost columns replicate the findings in the preceding paragraph. Turning to the viewpoint variables—as expected—forecasters who saw nuclear deterrence as fragile saw: (a) more junctures when the Cold War could have escalated into a thermonuclear war (*r*(34) = -.79); (b) higher risk of nuclear war 10/25/50/100 years out (*r*(34)'s = -.55, -.68, -.72 , -.77, *p*'s < .001). An unexpected result however was that those more skeptical of end-of-history arguments and confident that world politics would stay highly competitive saw less risk of nuclear war over the next century (*r*(34)'s = -.37, -.43, -.47, -.41, *p*'s <= .03).

Another unexpected result was that views on nuclear deterrence did not predict perceived risk of nuclear proliferation (centered NP forecasts), *r*(34) = .00. Nor did proliferation risk predict judged risk of nuclear war over the next century: *r*(34) = -.07, .03, .05, .15 (*p*'s > .3). One explanation for these disconnects is that worry about nuclear war is a joint function of seeing nuclear deterrence as fragile and seeing high proliferation risk. We lack the statistical power to test this interaction hypothesis but it is worth testing. Looking at the accuracy of combined 10- and 25-year nuclear-war projections (last row of Figure 6), the better forecasters saw nuclear deterrence as robust (*r*(34) = -.51); saw fewer nuclear-war close calls in the Cold War (*r*(34) = .43); and possessed greater proliferation expertise (*r*(34) = -.35).

Of most interest are the unresolved nuclear-war projections out to 2047 and 2097. Should we put more confidence in 50- and 100-year projections of forecasters who were: (a) more accurate on 10- and 25-year nuclear-war and nuclear-proliferation outcomes; (b) more scope-sensitive in estimating cumulative risk? Bayesians would find it surprising if there were zero diagnostic value in knowing the spread between elite and regular forecasters' judgments—and Skeptics would find it surprising if that value were appreciably above zero.

Proliferation experts satisfied requirements (a) and (b). Table 8 shows, relative to non-experts, they saw lower 10- and 25-year risks of nuclear war (10-year: *M* = 0.05 vs. *M* = 0.13; 25-year: *M* = 0.11 vs. *M* = 0.18) but saw similar 50- and 100-year levels of risk (50-year: *M* = 0.29 vs. *M* = 0.30; 100-year: *M* = 0.40 vs. *M* = 0.42). This convergence in 50- and 100-year forecasts means we cannot leverage knowledge of forecasting skill to reduce uncertainty about nuclear war by 2047 or 2097. But looking back in time, we can still make informative comparisons between the actual world with no nuclear war between 1997-2022 and two counterfactual worlds in which nuclear war erupts either between 1997-2007 or 2007-2022. In the real world, top proliferation forecasters had better accuracy scores on nuclear war in the 10- and 25-year ranges (Brier scores of NP experts vs non-experts: *M* = 0.01 vs. *M* = 0.06; log scores: *M* = 0.09 vs. *M* = 0.21). But in the counterfactual nuclear-war worlds, top forecasters take a reputational hit. And that hit is especially big when nuclear war occurs earlier—and we use a log scoring rule that is particularly punitive toward putting tiny probabilities on things that do happen. Assuming nuclear war between 1997 and 2007, the expert vs. non-expert contrasts are: *M* = 0.85 vs. *M* = 0.73; log scores: *M* = 3.21 vs *M* = 2.63. Assuming nuclear war between 2007 and 2022, the contrasts are *M* = 0.40 vs. *M* = 0.37; log scores: *M* = 1.45 vs. *M* = 1.29.

Here lies a cautionary tale: expect bolt-from-the-blue events to topple top forecasters whenever elite opinion was more dismissive of the risks than was mass opinion. This is a logical truth, not a contingent empirical proposition that can be overturned by the next study. If we want to incentivize forecasters not to miss cataclysms, we should switch from Brier scoring, appropriate for Gaussian distributions of routine outcomes, to logarithmic scoring appropriate for power-law distributions with extreme tail risk. Log scoring better aligns forecasters' and society's goals: a shared aversion to under-estimating existential threats.

Discussion

Meliorists can now claim systematic evidence for long-range geopolitical forecasting skill, an elusive phenomenon that some Skeptics had declared impossible (Taleb & Blyth, 2013) and one for which all previous evidence was anecdotal. Proliferation experts beat both well-educated generalists and base-rate extrapolation across time on the key empirical-accuracy indicator: they assigned higher probabilities when proliferation occurred—and lower values when it did not. Achieving a higher Hit rate at a lower False-Alarm rate proves proliferation experts were not indiscriminately crying wolf. Experts' edge even held across controversy and close-call-counterfactual challenges to accuracy scores, which blunts the flukiness-of-history objection. Moreover, proliferation experts did better on logical-coherence indicators. Their judgments were more scope sensitive and aligned with the normative model for compounding cumulative risk. And they did all of this under far-from-ideal conditions: making rapid-fire judgments, about one nation-state per minute. They drew on insights more accessible to epistemic-community insiders than to outsiders—a hallmark of genuine expertise.

A natural next question is: How much should Radical Skeptics change their minds? But that question is premature. The findings did not always break against them. Expertise failed to translate into accuracy on over half of the questions: those on border-change/secession. Moreover, the data are limited to a narrow slice of history—and the questions posed a deliberately biased sample from the universe of possible questions: slow-motion variables chosen to give forecasters a fighting chance. It is unwise to draw sweeping conclusions from so wobbly an evidentiary base. Whatever long-range predictability emerged is due to loading the methodological dice: posing easy questions in a placid period of history.

Each side is now armed with talking points: Meliorists, their proof-of-concept demonstration; Skeptics, their reasons for debunking it. One could call it a draw. But that too would be too facile. The problems run deeper than a debate over a dataset; the debate itself is flawed. Each school of thought has too many conceptual degrees of freedom for neutralizing disagreeable findings, enough to stalemate debates over virtually any dataset.

That is why we need an unusually long Discussion section that resets ground rules. Let's start by imagining two instructively extreme hypothetical datasets. In Counterfactual Dataset #1, experts fizzle. Their long-range forecasts are never better than non-experts and both groups lose to base-rate extrapolation across topics and time. In Counterfactual Dataset #2, experts beat non-experts and algorithms across the board. Skeptics resonate to Dataset #1; Meliorists, to Dataset #2.

We treat each school of thought as a research program in Lakatos's (1976) sense: grounded in hard-core assumptions buffered from reality by a protective belt of auxiliary hypotheses. The hard core of Radical Skepticism is a set of interlocking reasons for supposing long-range forecasting is futile: (a) the sensitivity of complex social systems to tiny tweaks of initial conditions that cause history to veer off course (butterfly effects); (b) the insensitivity of our measures and noisiness of the world combine to make it impossible to detect, less still predict, these tweaks; (c) once a world has veered off course, it is impossible to predict whether negative-feedback-loop equilibrium forces will bring it back on track or positive-feedback-loop forces will accelerate deviations; (d) any seemingly systematic variation in long-range foresight will therefore prove ephemeral and is most parsimoniously treated as a lucky streak.

Confronted by challenges to the hard core—like Dataset #2—defenders can draw on handy auxiliary hypotheses: (a) a moving-goalpost definition of what counts as a meaningful demonstration of long-range forecasting in a given domain or period; (b) the flexibility to shift standards of evidence and apply a "must-I-believe-this?" standard to inconvenient facts and a "can-I-believe-this?" standard to convenient ones. When pockets of predictability pop up, defenders write them off as aberrations and outliers.

The hard core of Meliorism is softer. Long-range forecasting is fallible but failure is not inevitable. This position rests on its own set of interlocking rationales: (a) complex systems are not as sensitive to tweaks of initial conditions as Skeptics posit; (b) negative-feedback-loop causality can create stable equilibria and pockets of predictability; (c) how long these pockets persist will depend on exogenous shocks that are themselves often somewhat predictable; (d) net predictability hinges on period of history, outcomes being forecast and skills of forecasters.

Confronted by challenges—like Dataset #1—these defenders can fall back on their own protective belt and attribute anomalies to: (a) picking domains where expertise does not yet confer knowledge of durable principles of causality but soon will; (b) relying on under-powered research designs; (c) failing to incentivize and train forecasters; (d) failing to recruit talented forecasters. Meliorists have as much flexibility as Skeptics to neutralize inconvenient facts.

To appreciate how fast each side can force a draw in debates, look at the questions each side leaves dangling. How much better would forecasters have to do to induce Radical Skeptics to concede ground? Would a longer series of bigger effect sizes suffice? Conversely, how much must top forecasters flop to move Meliorists? Would it require zero evidence of ability to beat extrapolation algorithms? The answers are murky because neither side has specified priors on the probability distributions of effect sizes or likelihood ratios on how much they would revise those priors in response to future data.

All this leaves each research program non-falsifiable. The scientific community need not though acquiesce to endless stalemates. It can require each school of thought to offer progressive-problem-shift defenses that not only patch up holes in explanations of existing phenomena but also advance *generative* hypotheses that stimulate new discoveries.

Our method of enforcing this norm is to incentivize rival programs to police each other under the aegis of Kahneman's model of adversarial collaboration (Kahneman, 2011; Mellers, Hertwig &

Kahneman, 2001). Creating the right incentives is a matter of setting the right conversational ground rules. Kahneman's original rules emphasize respectful perspective-taking and joint problem-solving. To that, we add a Bayesian emphasis on eliciting ex ante reputational bets and a Lakatosian emphasis on constraining ad hoc-ery:

> (a) Master the other side's perspective to the point where you can reproduce their arguments to their satisfaction. It is impossible to jointly design studies to test each other's hypotheses if one does not understand how the other side thinks;
>
> (b) Agree to constrain defenses against awkward data, which caps second guessing of features of jointly designed studies;
>
> (c) Make bold reputational bets that expose hard-core beliefs to risk, which requires specifying prior probability distributions of effect sizes and likelihood ratios that tell us how much one would change one's minds if agreed-on studies yielded varying effect sizes. Suppose Skeptics put a 10% chance on an effect size for expertise of >.3 on forecasts >25 years out and Meliorists see a 50% chance, 5X likelier than skeptics. A jointly designed study has the potential to have a big impact on the credibility of each camp.

To enforce rules of the game, we propose Lakatosian scorecards for holding each side accountable to peer reviewers who monitor: (a) how forcefully proponents dismiss dissonant findings; (b) how actively they explore progressive problem shifts that yield surprising discoveries. There is unavoidable subjectivity here. What one reviewer calls dogmatism, another might label, less pejoratively, Threat Deflection. What one reviewer calls Generativity, another might write off as a fishing expedition. Appraising scientific conduct is not noise-free. But the extremists will stand out. Failing grades—high Threat Deflection and low Generativity—go to those who offer no testable reasons for reneging on reputational bets.

We now put adversarial collaboration to work on improving debates over actual long-range data.

(a) Sampling Bias in Selecting Forecasting Questions. Radical Skeptics protest that Moderate Meliorists loaded the dice by focusing on slow-motion variables that gave long-range forecasters a fighting chance. Moderate Meliorists announced this plan from the outset so conceding this point is easy. But the next steps are harder. Each side must lay out their views more precisely than normal. Radical Skeptics might offer this reputational bet: expect the effect size of expertise in a well-designed study to be zero beyond 5 years, with a standard deviation of 0.1 (making the 0.4 effect size for proliferation expertise an extreme outlier). Meliorists might counter with their own wager: expect an effect size of .20, with a standard deviation of .15 (making 0.4 far less exceptional). The parties would then specify Bayesian likelihood ratios and commit to changing their minds as a function of who is closer to the mark.

All this would be progress but the devil lurks in details. The textbook solution to the sampling-bias problem is stratified random sampling from the universe of all questions bearing on high-stakes debates. But no one knows how even to define that universe so the parties will need to co-develop creative short-cuts. One work-around is to set up offsetting biases: each school nominates tough-but-fair challenges for the other. The parties will also need to co-develop a definition of expertise and baseline metrics that experts must beat. The more unqualified the baseline (non-expert) group, the easier it is to obtain big expertise effect sizes. The EPJ baseline was highly-educated generalists, but it is easy to deflate or inflate effect sizes by picking less or more sophisticated groups. There are many ways to load the dice.

Lakatosian Scorecard:

- Meliorists who reject even the sampling-bias complaint: high Threat Deflection, potentially disqualifying.
- Radical Skeptics and Meliorists who offer bold hypotheses (high-risk-of-falsification) and co-develop solutions to research-design challenges: high Generativity.

(b) Skill vs. Luck. Radical Skeptics trace proliferation experts' success to blind luck. Their preconceptions just happened to mesh with events at that phase of history. Expertise in 1997-2022 had less predictive value for border-change than proliferation because one time/topic combination made it easier for one set of experts to shine. Perhaps proliferation questions required no more knowledge than attentive readers could glean from the elite press. Anyone paying attention could quickly write off proliferation suspects such as South Africa, Brazil, Argentina, Nigeria, Sweden and Turkey. In this view, we found nothing surprising: proliferation experts were just doing their job of tracking proliferation news carefully and if we toss out the easy questions, with lots of forecasts close to zero, the experts' edge shrinks to zero.

Meliorists should push back by asking Skeptics for their precise operational definition of easy vs. hard cases. For it turns out proliferation expertise held up well across a range of definitions. Experts beat generalists with a restrictive definition that treats only the three actual proliferation incidents as hard—or if we expand the set to include Iran, the nation-state widely seen in 2022 as on the cusp of joining the nuclear club—or further expand the set to include other states of active interest to monitoring agencies in 1997—Iraq, Syria and Libya.

Radical Skeptics should concede an element of skill here but can still retrench round their trump argument: sampling bias. They can depict the proliferation-expert effect size as an outlier, an isolated pocket of predictability in one period of history. This concession contains damage to their hard core and is consistent with the rest of the dataset where expertise did not help.

Lakatosian Scorecard:

- Radical Skeptics who refuse to retreat from the softball-question objection: high Threat Deflection, potentially disqualifying.
- Radical Skeptics and Meliorists who work out a research design to reduce question-sampling bias in future studies: high Generativity.

(c) Domain Specificity of Expertise. Thus far, Skeptics have been on the defensive. Now it is the Meliorists' turn. How much should they concede in response to experts' failure to out-perform non-experts on border-change/secession? They have two defenses. The first is to invoke suboptimal working conditions: the hurried way in which the study elicited forecasts from ad hoc samples, hardly conducive to getting the best from the best. The second is to invoke differential scientific development across fields: relative to the nuclear proliferation domain, there were fewer relevant and durable principles of causality to guide border-change/secession forecasters. Each strength has strengths and weaknesses. Invoking poor working conditions is plausible because it connects to a broader cognitive debate over how much judgment improves as a function of time to think, training and other support (Chang et al., 2016). But poor work conditions should have handicapped both groups of forecasters, and it did not. So the burden of proof is on Meliorists to show that the science in one domain is "less developed" than in another. This defense is more plausible: the nuclear-engineering prerequisites for building atomic bombs flow from scientific principles much better defined than the political-science prerequisites for triggering shifts in national boundaries.

The next step is venturing guesses about effect sizes: How much would accuracy rise with better working conditions and forecasters? Research suggests effect sizes for forecaster training in the 8-12% range (Chang et al., 2016) and for superforecaster selection in the 25-40% range (Tetlock & Gardner, 2015)—though both lines of work involve short-range forecasting. Plus the unknown: how much will accuracy rise as a function of the soundness of the relevant science?

If they hold true to their hard-core convictions, Radical Skeptics will have the much easier effect-size estimation task. Their answers should be zero across the board.

Lakatosian Scorecard:

- Meliorists who invoke both working-conditions and differential-scientific-development explanations for weaker forecasting on border-change/secession and offer testable hypotheses: high Threat Deflection and high Generativity.
- Radical Skeptics who refuse to budge if accuracy improves under better working conditions or in sounder-science domains: high Threat Deflection and low Generativity.

(d) Logical Coherence and Empirical Accuracy. Radical Skeptics see the search for correlates of accuracy grounded in how forecasters think as futile: accuracy is a function of luck, fleeting by-products of fluky match-ups between forecasters' slow-moving priors and a fast-moving world. Meliorists disagree and cite the evidence at hand on links between logical coherence and accuracy indicators. Top forecasters made more scope-sensitive cumulative-risk judgments. They not only out-performed empirically, making more accurate initial 5-year probability estimates. They out-performed logically, adjusting initial estimates for longer periods in ways consistent with the axioms of probability theory. These data confront Skeptics with a trade-off.

Lakatosian Scorecard.

- Skeptics can treat the coherence-accuracy findings as a one-off that will not replicate—and score high on Threat Deflection but low on Generativity.
- Or they can score lower on Threat Deflection by conceding that logically coherent thinkers have an accuracy edge—and score higher on Generativity by working with Meliorists to identify when that edge is more or less pronounced.

Advancing Adversarial Collaborations Further. Skeptics' final retrenchment is to dismiss the current findings as a fluke and challenge Meliorists to show any generalizability beyond the topics, times and forecasters studied here. Meliorists could respond with either statistical or theoretical arguments. They can make a statistical case for limited generalizability by fitting log-logistic equations to decay rates in accuracy in the 25-year windows, 1988-2013 or 1997-2022, and projecting them out 50 and 100 years. The log-logistic model for these projections is defined on the domain ranging from time zero to infinity with $\alpha, \beta > 0$:

$$BrS(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^{-\beta}}$$

At time zero, the model takes on the value of zero (when we assume forecasters know the true state of the world) and has an upper bound of 1, allowing for convergence at any point along the spectrum of accuracy scores and for fitting either concave, convex, or mixed patterns of change.

Figure 7 shows the averaged question-level Brier scores (dots) and the 10%-90% interval of Brier-score ranges within each domain, with the best-fitting log-logistic model for each group of forecasters.[5] The advantage of expertise is visible from the differential elevation of log-logistic curves in the proliferation domain—and there is no such effect in the BCS domain. The dashed line represents the error for a naive "50/50" forecasting, with the linear models suggesting an advantage of expertise that projects even beyond fifty years for the proliferation domain. Indeed, proliferation experts did well enough that if we project the annual modeled decline in error from their 5-year forecasts (Brier = .02) through the 25-year forecasts (Brier = .05) and then to 50 or 100 years, these forecasters would still handily beat a 50-50% forecasting strategy. While non-experts in the proliferation domain still beat the 50-50% forecasting strategy over 100 years, their projected error increases more quickly than the experts'. Both experts and non-experts in the border-change domain are projected to match the 50-50% guessing strategy just after 50 years.

A strong Meliorist response will however require more than log-logistic extrapolations. We need a theory that treats the current cases as pieces of a larger puzzle. Proposing that theory is beyond our scope, but we can specify three factors that should determine the feasibility of predicting long-range predictability, each with testable hypotheses:

1. The 5-to-25-year data points suggest slower accuracy decay for forecasters who out-perform early on. Pinpointing the sources of this initial-period epistemic edge could improve our ability to predict longer-run predictability.

2. The stronger the historical cases for close-call counterfactuals or controversy adjustments, the faster the accuracy decay—and the likelier it is that logistic projections of long-range accuracy will be off due to violations of stationarity.

3. Accuracy decay will accelerate as a function of volatility in outcome base rates (e.g., a sudden surge of border-change events). Top forecasters made judgments that were closer to the actuarial-risk compounding model and thus more vulnerable to volatility, which raises the possibility of a trade-off between maximizing accuracy in shorter and longer-range tournaments. Were top forecasters niche-adapted to quasi-normal distributions of outcomes and thus at more risk of being blind-sided by power-law probability distributions?

[5] The best-fitting models were found by minimizing the squared error between the model and the averaged question Brier scores (*n* = 126 for NP domain, *n* = 120 for BCS domain).

A natural next step for Meliorists is to identify the modes of reasoning by which some forecasters achieve an early epistemic edge. Here researchers have already made some progress. Shorter-range studies have found better forecasters offer more dialectically complex rationales (Karvetski et al., 2021; Tetlock, 2005), are better Bayesian updaters (Mellers et al., 2015), and distinguish more degrees of uncertainty (Friedman et al., 2017).

Such findings suggest early epistemic edges are more than luck—and the winning modes of thinking are not mysterious. For instance, the cognitive-science literature on the power of simple models to reproduce experts' judgments in diverse domains (Dawes et al., 1989) implies that experts can achieve early epistemic edges in surprisingly straightforward ways. To predict proliferation predictions, we may only need forecasters' answers to a handful of questions: (a) does entity x has an active weapons program?; (b) does it perceive an existential threat?; (c) can its security concerns be allayed by reassurance-building steps, like alliances and treaties?; (d) can it be induced to give up its program by negative pressure (sanctions/military strikes) or positive pressure (economic-technical help)? Answers to (a) and (b) of "yes" and answers to (c) and (d) of "no" imply elevated risk. To paraphrase Dawes' summary of policy-capturing work, you just need to know what variables to look at—and how to add.

Although experts dislike Dawesian reductionism, there is no contradiction between two claims: expert judgment is often simpler than many experts like to think, and some experts are better than others at cleaving nature at its joints. Expertise is less about computational crunching power than about framing incisive questions. Consider a prescient rationale for a 1997 set of forecasts: "many states can build the bomb but powerful forces deter them. I worry about South Asia and North Korea. International sanctions are no match for nationalist passions in those regions. India and Pakistan are on the edge of declaring. North Korea is 5 to 10 years away. For me, the Middle East is the unknown: Iran, Iraq and even Libya and Syria have ways and means. But the Israelis and USA are watching—and won't make bomb construction easy." Judging from the pattern of forecasts, this rationale was common knowledge among proliferation experts but not outsiders. And it detracts nothing if a simple model can capture how experts pulled off the signal detection feat of a high Hit and low False-Alarm rate. Threading the needle between excessive optimism and pessimism is impressive, even after policy-capturing demystifies it by revealing the underlying formula: a mix of region-specific insights and transportable knowledge about game theory, human nature, and nuclear engineering.

Researchers can also foster adversarial collaborations among forecasters and encourage them to share constructive critiques of each other's rationales. One index of "constructive" could be based on Bayesian networks of if-then propositions underlying forecasts (Carless et al., 2021). Suppose one forecaster in 1997 saw a 90% chance of North Korea becoming a nuclear power in 10 years if it survived but only a 20% chance of survival—and another made mirror-image judgments. Each arrives at an 18% answer of a nuclear-armed North Korea by 2007. Adversarial collaborations could increase their odds of correcting each other's mistakes. And Bayesian models could increase researchers' odds of capturing this process and figuring out who is right for the right reasons (sound reasons yielding accurate predictions), right for the wrong reasons (offsetting errors producing accurate predictions), wrong in their predictions but sound in their rationales, or wrong in both their predictions and reasoning.

Learning from Our Mistakes. A casual reader of a research-methods text could easily enumerate the flaws in our small exploratory study—and ways of fixing them. We mention three fixes here. First, follow-up work should replace the 11-point probability scale with a continuous measure, like Karger et al.'s (2021) nonparametric probability scaling which elicits paired judgments: (i) will a nuclear weapon ever be used in combat?; (ii) if yes, by what date will that event become 25%/50%/75%/99%... likely to have happened? This avoids the distortions from forcing forecasters to choose between 0 and .1. Forecasters with granular assessments of risk—often the best ones (Friedman et al., 2017)—can offer judgments as infinitesimally close to zero as they see fit. Second, although there is no perfect solution to sampling bias in question selection, there is a good solution. Adversarial collaborators can create offsetting sources of sampling bias by taking turns nominating questions to foil the other side. Third, follow-up work could give forecasters even more of a fighting chance. Eliciting anonymized best guesses, as was done here, is just a start. It defines a baseline for gauging the accuracy boosts we can get from better-designed incentives, training, and team interaction (Tetlock & Gardner, 2015).

The most ambitious extension of the current work would address its deepest shortcoming: the impoverished accuracy-feedback forecasters get from a single run of history. We can only assess accuracy across forecasts—and never know whether the true annual risk of nuclear war in 1997-2022 was 5% (and we were lucky) or a billion times smaller. For ground-truth probabilities of events, we must turn to simulations of complex systems, like battlefields or financial markets, that give forecasters fast feedback on how adept they are at distinguishing negative-feedback-loops that oscillate in equilibrium zones and produce Gaussian distributions of outcomes from positive-feedback-loop dynamics that produce power-law distributions of outcomes. If the latter, the next world war might claim 10X or 100X the lives lost in World War 2.

Cumulative Conversations. Meliorists believe they can leverage knowledge of forecaster track records to answer once intractable questions—that if they knew forecaster track records over the last 25 or 50 years, they could better estimate threats and opportunities over the next 50 to 100 years. Skeptics are wary. All inductive reasoning is tentative; past performance is no guarantee of future performance. Meliorists understand the wariness but respond that giving zero weight to top forecasters and algorithms implies absolute certainty that an abrupt discontinuity has neutralized all signal value from the past. And absolutists tend to be poor forecasters (Tetlock, 2005). Skeptics do not deny that, all else equal, we are better off when well-intentioned policymakers have well-calibrated estimates of probable consequences of options. But all else is rarely equal. They worry more about quixotic research programs that offer false reassurance that the unknowable is knowable (Kay & King, 2020) than they do about missing chances to enhance foresight (Tetlock, Lu, & Mellers, 2022). Counterfactual adjustments to accuracy scores remind us that top forecasters are one ill-timed nuclear war from being knocked off their pedestal.

Adversarial collaborations can accelerate how fast these sorts of conversations can converge on the truth by incentivizing each school of thought to refine its position to address the most compelling objections of the other school. Of course, that does not guarantee perfect Bayesian convergence. Schools of thought may disagree not only over facts but also values that set their thresholds of proof for judging facts. But partial convergence is often within reach.

References

Ashton, A. H., & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. Management Science, 31(12), 1499-1508.

Bas, M. A., & Coe, A. J. (2016). A Dynamic Theory of Nuclear Proliferation and Preventive War. In International Organization (Vol. 70, Issue 4, pp. 655–685). Cambridge University Press (CUP). https://doi.org/10.1017/s0020818316000230

Bell, M. S. (2015). Examining Explanations for Nuclear Proliferation. In International Studies Quarterly (Vol. 60, Issue 3, pp. 520–529). Oxford University Press (OUP). https://doi.org/10.1093/isq/sqv007

Black, C. (2004). Types of secessionist conflict : explaining the conditions that affect the duration and violence of secessionist conflicts [UNSW Sydney]. https://doi.org/10.26190/UNSWORKS/6726

Bolger, F. & Wright, G. (1994). Assessing the quality of expert judgment: Issues and analysis. *Decision Support Systems, 11,* 1-24.

Carless, T. S., Redus, K., & Dryden, R. (2021). Estimating nuclear proliferation and security risks in emerging markets using Bayesian Belief Networks. In Energy Policy (Vol. 159, p. 112549). Elsevier BV. https://doi.org/10.1016/j.enpol.2021.112549

Carter, D. B., & Poast, P. (2016). Why Do States Build Walls? Political Economy, Security, and Border Stability. In Journal of Conflict Resolution (Vol. 61, Issue 2, pp. 239–270). SAGE Publications. https://doi.org/10.1177/0022002715596776

Cimbala, S. J. (2017). Nuclear Proliferation in the Twenty-First Century: Realism, Rationality, or Uncertainty? Strategic Studies Quarterly, 11(1), 129–146. http://www.jstor.org/stable/26271593

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668-1674.

Deutsch, D. (1998). *The fabric of reality*. Penguin UK.

Dietrich, F. (2010). Bayesian group belief. *Social choice and welfare*, *35*(4), 595-626.

Friedman, J., Baker, J., Mellers, B. A., Zeckhauser, R., & Tetlock, P. E. (2017). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly*, 62(2), 410-422.

Fukuyama, Francis (1989). The End of History? *The National Interest* (16): 3–18.

Fukuyama, F. (1995). Reflections on the end of history, five years later. *History and theory*, 27-43.

Gaba, A., Popescu, D. G., & Chen, Z. (2019). Assessing uncertainty from point forecasts. Management Science, 65(1), 90-106.

Gaddis, J. L. (1992). International relations theory and the end of the Cold War. *International security*, *17*(3), 5-58.

Gaddis, J. L. (2019). *On grand strategy*. Penguin.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650.

Gleditsch, K. S., & Ward, M. D. (2013). Forecasting is difficult, especially about the future. In Journal of Peace Research (Vol. 50, Issue 1, pp. 17–31). SAGE Publications. https://doi.org/10.1177/0022343312449033

Gleick, J. (2008). *Chaos: Making a new science*. Penguin UK

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359-378.

Goemans, H. E., & Schultz, K. A. (2016). The Politics of Territorial Claims: A Geospatial Approach Applied to Africa. In International Organization (Vol. 71, Issue 1, pp. 31–64). Cambridge University Press (CUP). https://doi.org/10.1017/s0020818316000254

Goetzmann, W. N., Kim, D., & Shiller, R. J. (2022). *Crash Narratives* (No. w30195). National Bureau of Economic Research.

Gove, M. (2016). Experts Like Carney Must Curb Their Arrogance. The Times, October 21, 2016.

Grundmann, R. (2022). Making sense of expertise. London: Routledge.

Huth, P. K. (1996). Standing your ground: Territorial disputes and international conflict. The University of Michigan Press.

Jervis, R. (1998). *System effects: Complexity in political and social life*. Princeton University Press.

Jervis, R. (2010). Why intelligence fails. In *Why Intelligence Fails*. Cornell university press.

Kahneman, D. (2011). Thinking, fast and slow. Farrar: New York.

Kahneman, D., Sibony, O., & Sunstein, C. (2021). Noise. New York: Farrar, Straus & Giroux.

Kaplow, J. M., & Gartzke, E. (2016). Predicting Proliferation: High Reliability Forecasting Models of Nuclear Proliferation as a Policy and Analytical Aid. [Calhoun: The NPS Institutional Archive].

Karger, E., Atanasov, P. D., & Tetlock, P. (2022). Improving Judgments of Existential Risk: Better Forecasts, Questions, Explanations, Policies. Oxford University: Future of Humanity Institute.

Kay, J. A., & King, M. A. (2020). *Radical uncertainty*. *Decision-making beyond the numbers*. Bridge Street Press.

Keren, G. (1987). Facing uncertainty in the game of Bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 39*, 98-114.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?* (pp. 205-259). Springer, Dordrecht.

Levinson, M. (2017). The Origins of Secessionist Conflict: Predicting When Governments Will and Will Not Permit Secession [The University of North Carolina at Chapel Hill University Libraries]. https://doi.org/10.17615/TW14-SS76

McCloskey, D. N. (1991). History, Differential Equations, and the Problem of Narration. *History and Theory*, *30*(1), 21–36. https://doi.org/10.2307/2505289

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. (2017). How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making, 12*(4), 369-380.

Mitchell, P. G. & Tetlock, P. E. (2023). Are progressives in denial About progress? Yes, but so is almost everyone else. *Clinical Psychological Science*.

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600.

Nisbett, R. & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological review, 84(3), 231.

Palley, A.B. and Satopää, V.A., (2023). Boosting the wisdom of crowds within a single judgment problem: Weighted averaging based on peer predictions. Management Science.

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. A. (2022a). Decomposing the effects of crowd-wisdom aggregators: The bias-information-noise (BIN) model. *International Journal of Forecasting*.

Satopää, V. A. (2022b). Regularized aggregation of one-off probability predictions. *Operations Research*.

Satopää, V. A., Salikhov, M., Tetlock, P. E., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *Management Science, 67*(12), 7599-7618.

Scoblic, J. P., Karvetski, C., & Tetlock, P. E. (2021). Did Sino-American relations have to deteriorate? A better way of doing counterfactual thought experiments. Retrieved from https://warontherocks.com/2021/07/did-sino-american-relations-have-to-deteriorate-a-better-way-of-doing-counterfactual-thought-experiments/

Solingen, E. (2018). Nuclear Proliferation: The Risks of Prediction. In SSRN Electronic Journal. Elsevier BV. https://doi.org/10.2139/ssrn.3275670

Stevens, J. (2012). Political scientists are lousy forecasters. New York Times, June 24, 2012, SR6.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Taleb, N. N., & Blyth, M. (2011). The black swan of Cairo: How suppressing volatility makes the world less predictable and more dangerous. *Foreign Affairs*, 33-39.

Tetlock, P. E., & Belkin, A. (Eds.) (1996). *Counterfactual thought experiments in world politics*. Princeton, NJ Princeton: Princeton University 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*, 639-652.

Tetlock, P. E. (2005, first edition; 2017, second edition). *Expert political judgment: How good is it? How can we know*? Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown.

Tetlock, P. E., Lebow, R. N., & Parker, G. (Eds.) (2006). *Unmaking the west: What-if scenarios that rewrite world history.* Ann Arbor, MI: University of Michigan Press.

Tetlock, P.E., Lu, Y., & Mellers, B. A. (2022). False dichotomy alert: Cultivating talent at probability estimation versus raising awareness of systemic risk. *International Journal of Forecasting*.

Yaniv I., Yates, J.F., Smith, J.E.K. (1991) Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin*, 110(3), 611–617.

Table 1

Summary of Data Collected in 1988 and 1997

Year	Sample Size			Viewpoint Variables	Forecasting Questions		
	Total	NP Specialists	BCS Specialists		NP Forecasts	BCS Forecasts	Nuclear War Forecasts
1988	13	4	--	No	Yes	No	No
1997	34	11	9	Yes	Yes	Yes	Yes

Note. NP is abbreviation for nuclear proliferation; BCS is abbreviation for border change/secession.

Table 2

Accuracy Scores for Nuclear Proliferation (NP) Forecasts

Timeframe	Baserate	Brier Score					Log Score	
		Predict Baserate	Predict No Change	All Forecasters	NP Non-Expert	NP Expert	NP Non-Expert	NP Expert
5	.02	.02	.02	.03	.03	.02	.11	.07
10	.10	.09	.10	.05	.06	.03	.18	.11
25	.12	.10	.12	.07	.08	.05	.26	.17
All	.08	.07	.08	.05	.06	.03	.18	.12

Note. For Log Scores, we set probabilities of 0 at .01 and probabilities of 1 at .99.

Table 3
Average Forecasts for Nuclear Proliferation (NP) Across Three Periods

Timeframe	NP Outcome	NP Average Forecast	
		Non-Expert	Expert
5	0	.05 (*n* = 617, *SD* = 0.12)	.03 (*n* = 281, *SD* = 0.11)
	1	.38 (*n* = 23, *SD* = 0.20)	.58 (*n* = 11, *SD* = 0.28)
10	0	.06 (*n* = 576, *SD* = 0.13)	.04 (*n* = 262, *SD* = 0.10)
	1	.44 (*n* = 64, *SD* = 0.23)	.62 (*n* = 30, *SD* = 0.26)
25	0	.13 (*n* = 567, *SD* = 0.20)	.09 (*n* = 258, *SD* = 0.18)
	1	.52 (*n* = 73, *SD* = 0.24)	.72 (*n* = 34, *SD* = 0.25)

Table 4

Accuracy Scores for Border Change/Secession (BCS) Forecasts

Timeframe	Baserate	Brier Score					Log Score	
		Predict Baserate	Predict No Change	All Forecasters	BCS Non-Expert	BCS Expert	BCS Non-Expert	BCS Expert
5	.10	.09	.10	.06	.06	.06	.21	.23
10	.13	.11	.13	.07	.07	.07	.25	.27
25	.23	.17	.23	.15	.15	.15	.51	.51
All	.15	.13	.15	.09	.09	.10	.33	.34

Note. For Log Scores, we set probabilities of 0 at .01 and probabilities of 1 at .99.

Table 5
Average Forecasts for Border Change/Secession (BCS) Across Three Periods

Timeframe	BCS Outcome	BCS Average Forecast	
		Non-Expert	Expert
5	0	.04 (*n* = 792, *SD* = 0.11)	.03 (*n* = 324, *SD* = 0.08)
	1	.37 (*n* = 88, *SD* = 0.26)	.28 (*n* = 36, *SD* = 0.23)
10	0	.07 (*n* = 770, *SD* = 0.14)	.06 (*n* = 315, *SD* = 0.12)
	1	.46 (*n* = 110, *SD* = 0.29)	.37 (*n* = 45, *SD* = 0.28)
25	0	.14 (*n* = 682, *SD* = 0.20)	.14 (*n* = 279, *SD* = 0.19)
	1	.41 (*n* = 198, *SD* = 0.33)	.37 (*n* = 81, *SD* = 0.29)

Table 6
Systematic Upward (Positive) or Downward (Negative) Bias in the Nuclear Proliferation (NP) Predictions

Timeframe	All Forecasters	Experts	Non-Experts
5	.02 [.01, .03]	.01 [-.01, .03]	.03 [.01, .04]
10	.00 [-.02, .01]	-.01 [-.05, .02]	.00 [-.02, .02]
25	.06 [.03, .08]	.04 [.00, .08]	.06 [.04, .09]
All	.02 [.02, .03]	.01 [.00, .03]	.03 [.02, .04]

Note. Values in brackets represent 95% bootstrap confidence intervals.

Table 7

Systematic Upward (Positive) or Downward (Negative) Bias in the Border Change/Secession (BCS) Predictions

Timeframe	All Forecasters	Experts	Non-Experts
5	-.03 [-.05, -.02]	-.04 [-.07, -.02]	-.03[-.05, -.01]
10	-.01 [-.03, .00]	-.02 [-.05, .00]	-.01[-.02, .01]
25	-.03 [-.05, -.01]	-.04 [-.08, .00]	-.03 [-.05, .00]
All	-.02 [-.03, -.01]	-.03 [-.05, -.02]	-.02 [-.03, .-01]

Note. Values in brackets represent 95% bootstrap confidence intervals.

Table 8

Averages (and Standard Deviations) for Expertise, Centered Forecasts, Normalized Brier Scores, Viewpoint Variables, and Nuclear War Forecasts

	NP Non-Expert	NP Expert	*p*-Value
Self-Assessed NP Expertise	2.13 (0.81)	4.45 (0.52)	.00
Self-Assessed BCS Expertise	2.78 (1.13)	3.27 (1.10)	.24
Centered NP Forecast	0.01 (0.03)	-0.01 (0.02)	.16
Normalized NP Brier Score	0.13 (0.28)	-0.19 (0.26)	.00
Fragility/Robustness	4.74 (2.34)	4.82 (2.04)	.92
How Often	2.04 (0.98	1.91 (0.94)	.70
Close Call	5.61 (1.56)	5.91 (2.07)	.68
Continuity	5.57 (1.83)	5.82 (1.83)	.71
Forecast Nuclear War 10yr	0.13 (0.17)	0.05 (0.05)	.03
Forecast Nuclear War 25yr	0.18 (0.19)	0.11 (0.10	.16
Forecast Nuclear War 50yr	0.30 (0.24)	0.29 (0.27)	.89
Forecast Nuclear War 100yr	0.42 (0.29)	0.40 (0.32)	.85
Brier Score Nuclear War 10yr	0.04 (0.08)	0.00 (0.01)	.03
Brier Score Nuclear War 25yr	0.07 (0.12)	0.02 (0.03)	.09

Note. Self-Assessed Expertise was rated on a 5-point scale.

Figure 1

Nuclear Proliferation (NP) Counterfactual Scenarios

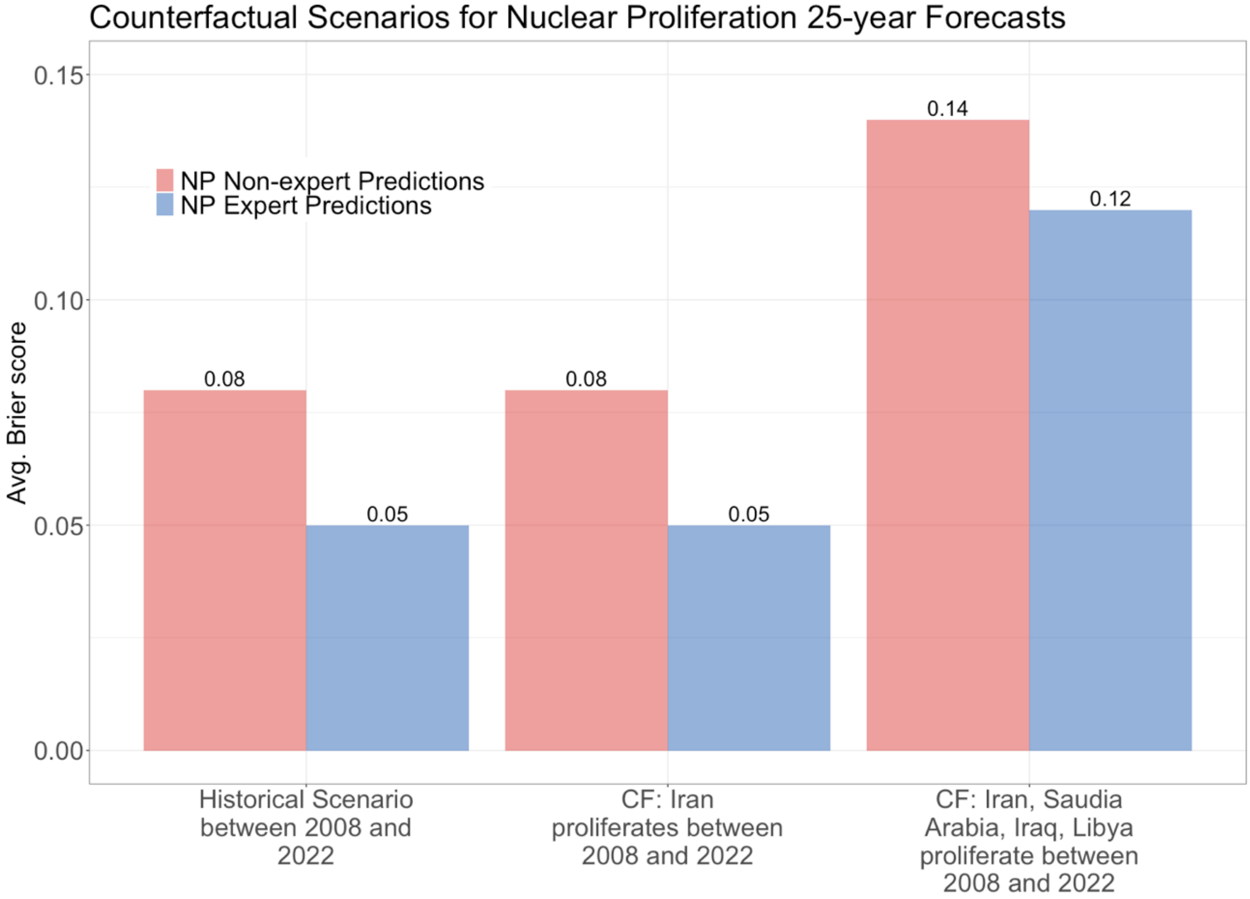


Figure 2

Border Change/Secession (BCS) Counterfactual Scenarios

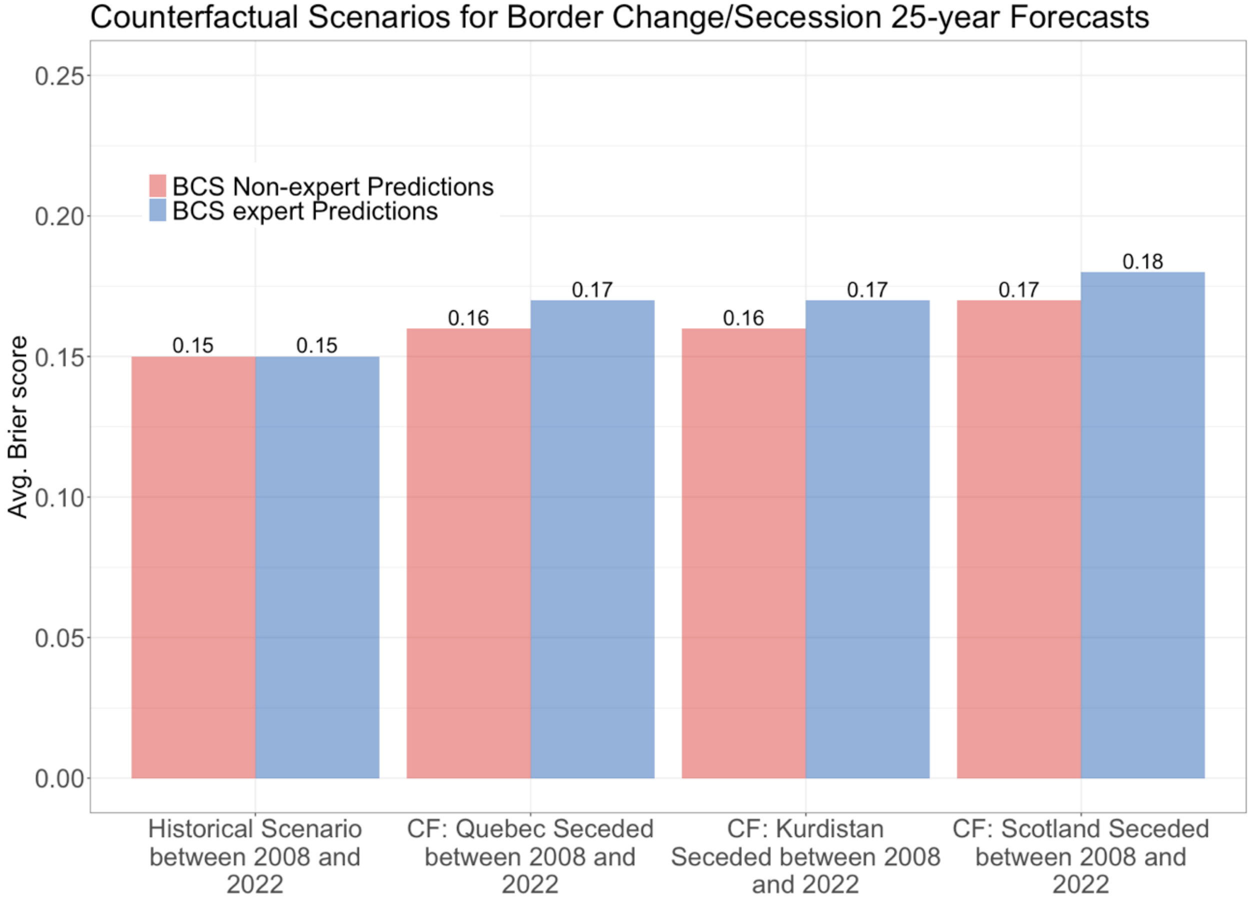
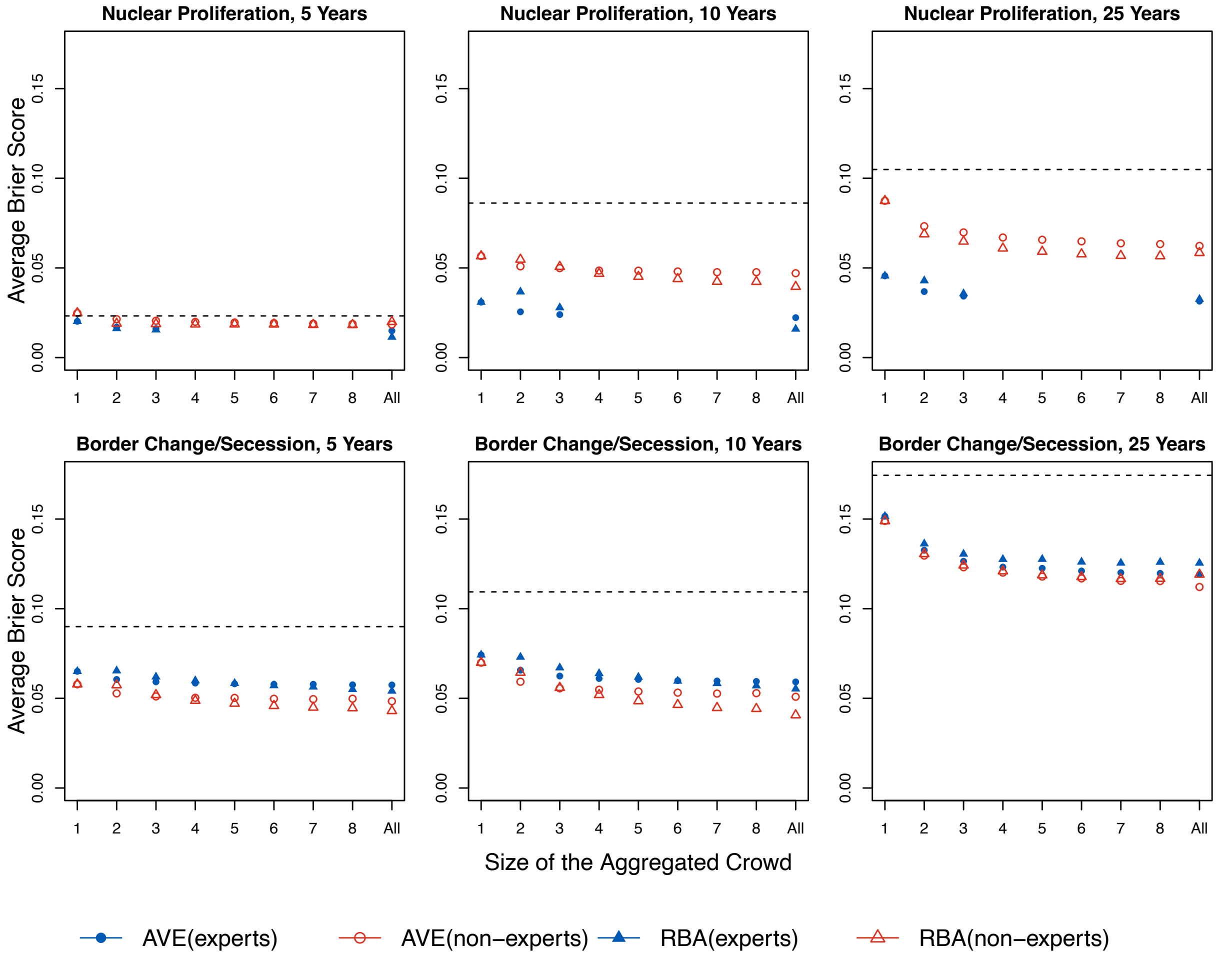


Figure 3

Brier Scores of the Simple Average (AVE), Regularized Bayesian Aggregator (RBA), and Base-Rate Extrapolation



Note. The dashed horizontal line represents base-rate extrapolation. The left-most points (above the label "1") represent scores from the average individual forecaster, and the right-most points (above the label "All") represent scores from aggregating all non-experts' or experts' predictions in our dataset.

Figure 4

Percentage of Forecasters Out-Performed by Three Aggregators: Simple Average (AVE), Regularized Bayesian Aggregator (RBA), and Base-Rate Extrapolation

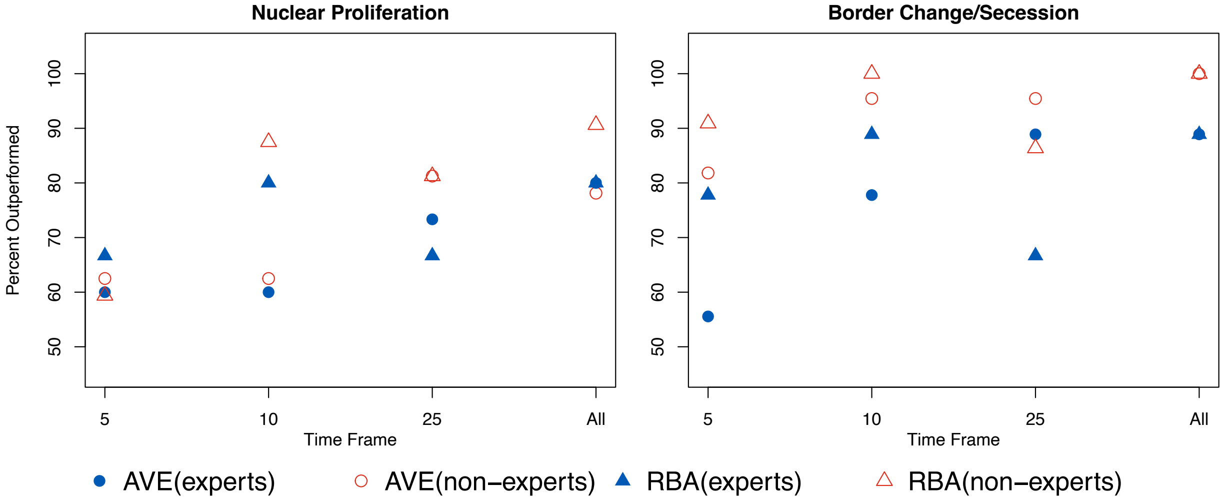


Figure 5

Average Differences Between Model-Derived Probabilities and Forecasters' Probabilities

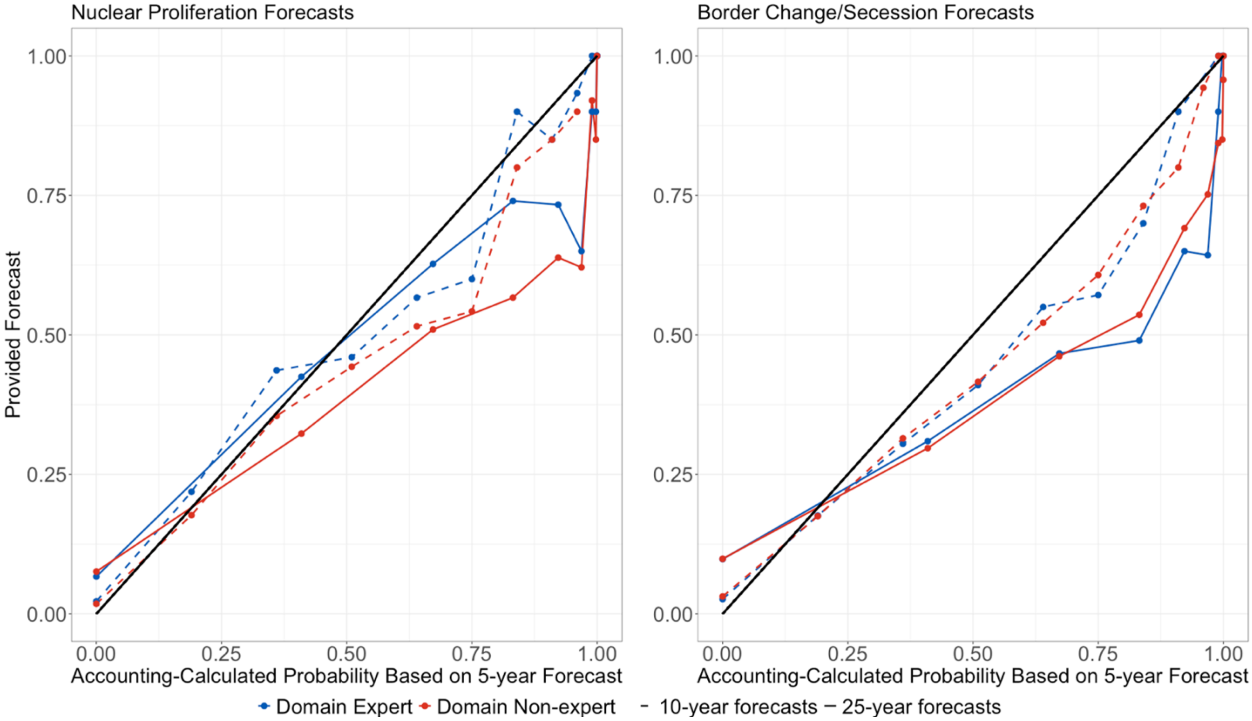
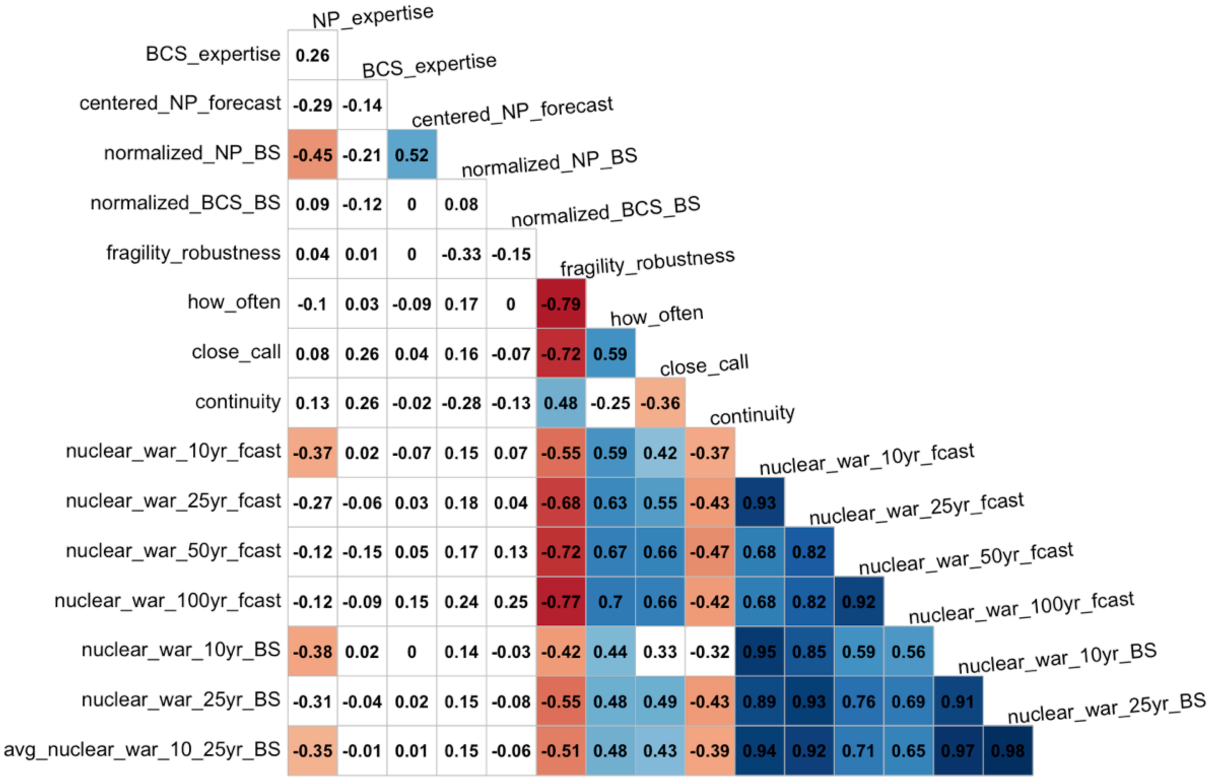


Figure 6

Correlations Between Expertise, Centered Forecasts, Normalized Brier Scores, Viewpoint Variables, and Nuclear War Forecasts

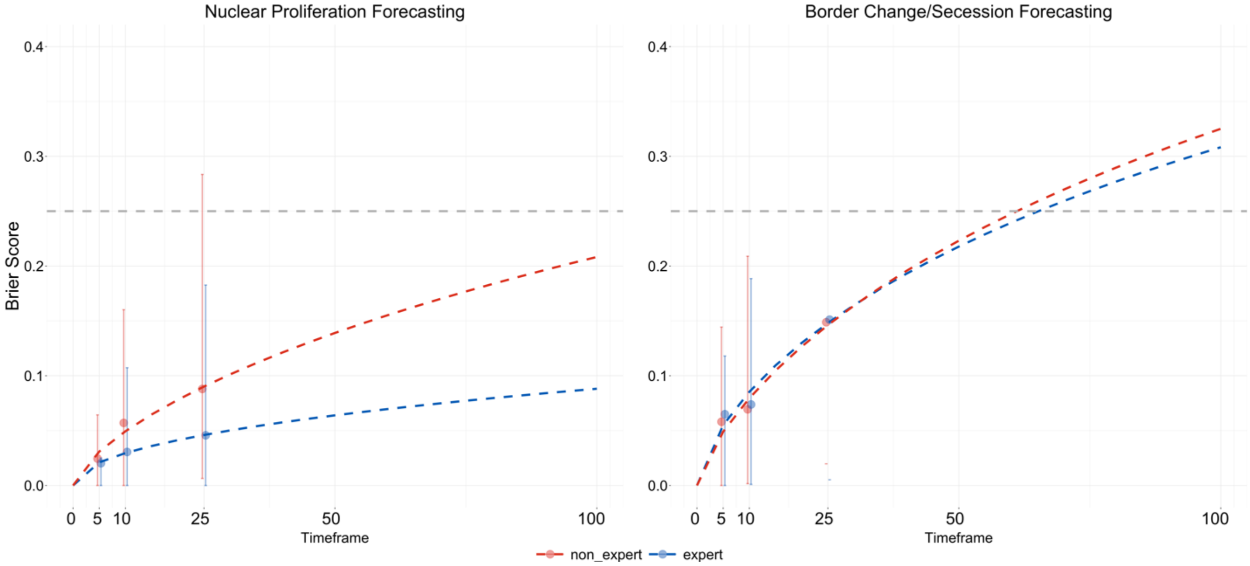


	NP_expertise	BCS_expertise	centered_NP_forecast	normalized_NP_BS	normalized_BCS_BS	fragility_robustness	how_often	close_call	continuity	nuclear_war_10yr_fcast	nuclear_war_25yr_fcast	nuclear_war_50yr_fcast	nuclear_war_100yr_fcast	nuclear_war_10yr_BS	nuclear_war_25yr_BS
BCS_expertise	0.26														
centered_NP_forecast	-0.29	-0.14													
normalized_NP_BS	-0.45	-0.21	0.52												
normalized_BCS_BS	0.09	-0.12	0	0.08											
fragility_robustness	0.04	0.01	0	-0.33	-0.15										
how_often	-0.1	0.03	-0.09	0.17	0	-0.79									
close_call	0.08	0.26	0.04	0.16	-0.07	-0.72	0.59								
continuity	0.13	0.26	-0.02	-0.28	-0.13	0.48	-0.25	-0.36							
nuclear_war_10yr_fcast	-0.37	0.02	-0.07	0.15	0.07	-0.55	0.59	0.42	-0.37						
nuclear_war_25yr_fcast	-0.27	-0.06	0.03	0.18	0.04	-0.68	0.63	0.55	-0.43	0.93					
nuclear_war_50yr_fcast	-0.12	-0.15	0.05	0.17	0.13	-0.72	0.67	0.66	-0.47	0.68	0.82				
nuclear_war_100yr_fcast	-0.12	-0.09	0.15	0.24	0.25	-0.77	0.7	0.66	-0.42	0.68	0.82	0.92			
nuclear_war_10yr_BS	-0.38	0.02	0	0.14	-0.03	-0.42	0.44	0.33	-0.32	0.95	0.85	0.59	0.56		
nuclear_war_25yr_BS	-0.31	-0.04	0.02	0.15	-0.08	-0.55	0.48	0.49	-0.43	0.89	0.93	0.76	0.69	0.91	
avg_nuclear_war_10_25yr_BS	-0.35	-0.01	0.01	0.15	-0.06	-0.51	0.48	0.43	-0.39	0.94	0.92	0.71	0.65	0.97	0.98

Note. Self-Assessed Expertise was rated on a 5-point scale. Shaded cells represent significance at the .05 level.

Figure 7

Log-Logistic Extrapolation of Brier Score for Both Domains


Nuclear Proliferation Forecasting
Border Change/Secession Forecasting
Brier Score
Timeframe
non_expert
expert