DOI: 10.1002/ffo2.173

COMMENTARY

WILEY

Exploring the limits on Meliorism: A commentary on Tetlock et al. (2023)

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

[1]Wharton School of the University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Forecasting Research Institute, Philadelphia, Pennsylvania, USA

[3]Good Judgment Inc, New York City, New York, USA

[4]INSEAD, Fontainebleau, France

Correspondence: Philip E. Tetlock, Wharton School of the University of Pennsylvania, Philadelphia, PA, USA.
Email: tetlock@wharton.upenn.edu

Funding information
Founders Pledge; Open Philanthropy

1 | EXPLORING THE LIMITS ON MELIORISM

The target article for this symposium, Tetlock et al. (2023), undercuts two oft-heard generalizations about geopolitical forecasting: namely, that specialists in world politics are hard-pressed to beat well-informed generalists (Mellers et al., 2015; Tetlock, 2017)—and that accuracy falls off the further into the future either specialists or generalists try to see (Satopää et al., 2021; Tetlock, 2017). The surprise finding was the superior performance of experts on nuclear proliferation in separating cases of proliferation and non-proliferation as far out as 25 years.

Our three commentators, all prominent geopolitical experts in their own right, highlight fruitful directions for follow-up work that can gauge whether the current findings are either: (a) one-off flukes (make enough specialist-generalist comparisons on enough topics on enough occasions, the laws of chance guarantee occasional wins for the experts), or (b) evidence of robust boundary conditions under which expertise is likely to translate into predictive skill.

All three commentators warn us, in their distinctive ways, to be wary of sweeping claims. Treisman (2023) pinpoints the pivotal issues: What makes something easier or harder to predict in the shorter or longer term? And when does expertise help? He proposes "causal regularity" as the key moderator of the feasibility of long-range forecasting and offers an intriguing example from his own research program of long-range forecasting proving easier than short-range: the impact of economic development on democratization. He also argues convincingly that expertise is likelier to add predictive value when the regularity of underlying causal processes falls in a sweet-spot zone: neither too regular (in which case, even non-experts get it right) nor too idiosyncratic and irregular (in which case, no one gets it right).

Treisman does not just advance hypotheses well worth testing in future work; he applies his framework to past work to explain why expertise was likelier to add predictive value in the nuclear proliferation domain than in the border-change domain. The drivers of border change are just too varied and case-specific, whereas nuclear proliferation is scientifically neater, lending itself to the enumeration of well-defined necessary and sufficient conditions.

Treisman's incisive commentary suggests to us the value of re-examining the data through the lens of signal detection theory. Figures 1 and 2 show receiver operating characteristic (ROC) curves of forecasts from specialists and generalists over time. Consumers of forecasts, the decisionmakers, should pay attention: each point on the ROC curves corresponds to a forecast probability and shows the tradeoff between the true-detection proportion and the false-positive proportion that would result if a decision-maker chose to act whenever the probability of occurrence exceeded the designated probability. So, (0.04, 0.82) on the experts' blue curve in the 5-year panel implies that if a decision-maker acted when proliferation experts' 5-year forecast probability reached 0.3, the decision-maker would catch 82% of all 5-year proliferation events (true positives) at the cost of a 4% false-positive rate. ROC curves track the co-occurrence of true and false positives as threshold probabilities move from 0 to 1.

ROC curves offer two ways to quantify performance. First, we can say experts dominate non-experts whenever their ROC curve always hovers above the nonexperts' curve. This holds true for all

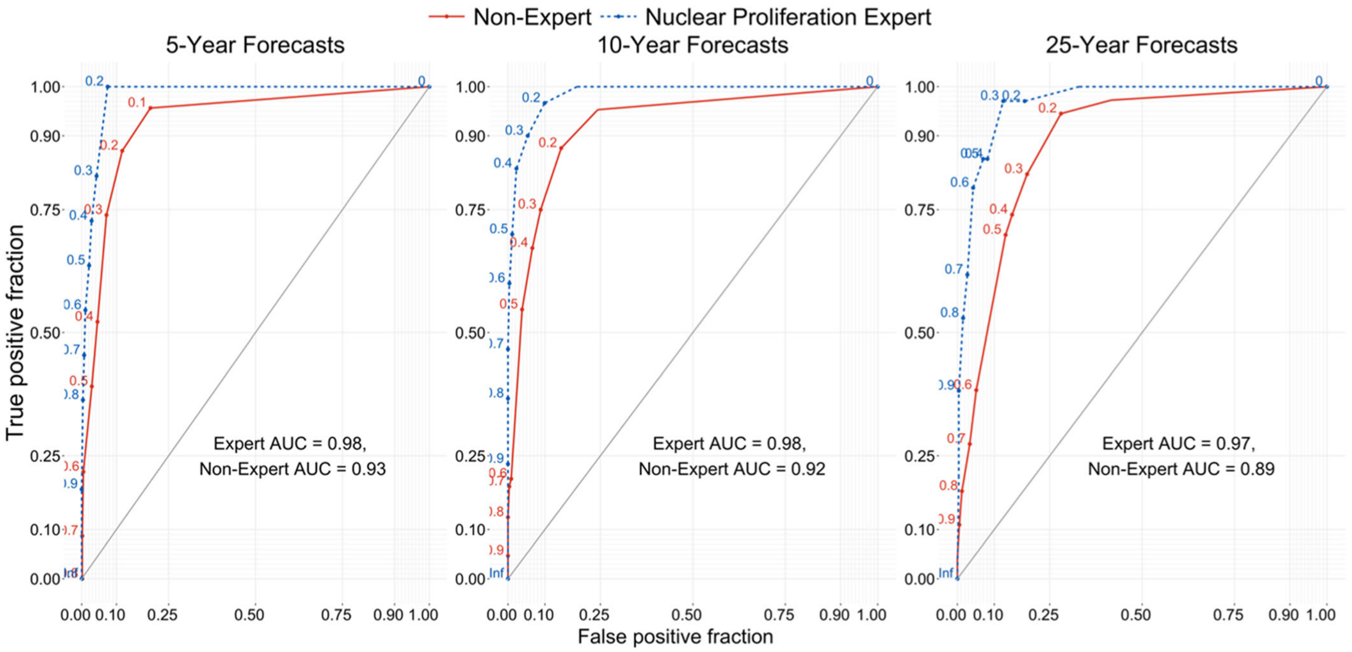

FIGURE 1 ROC analysis of nuclear proliferation forecasts of experts and non-experts across 5, 10, and 25 years. AUC, area under the curve; ROC, receiver operating characteristic.

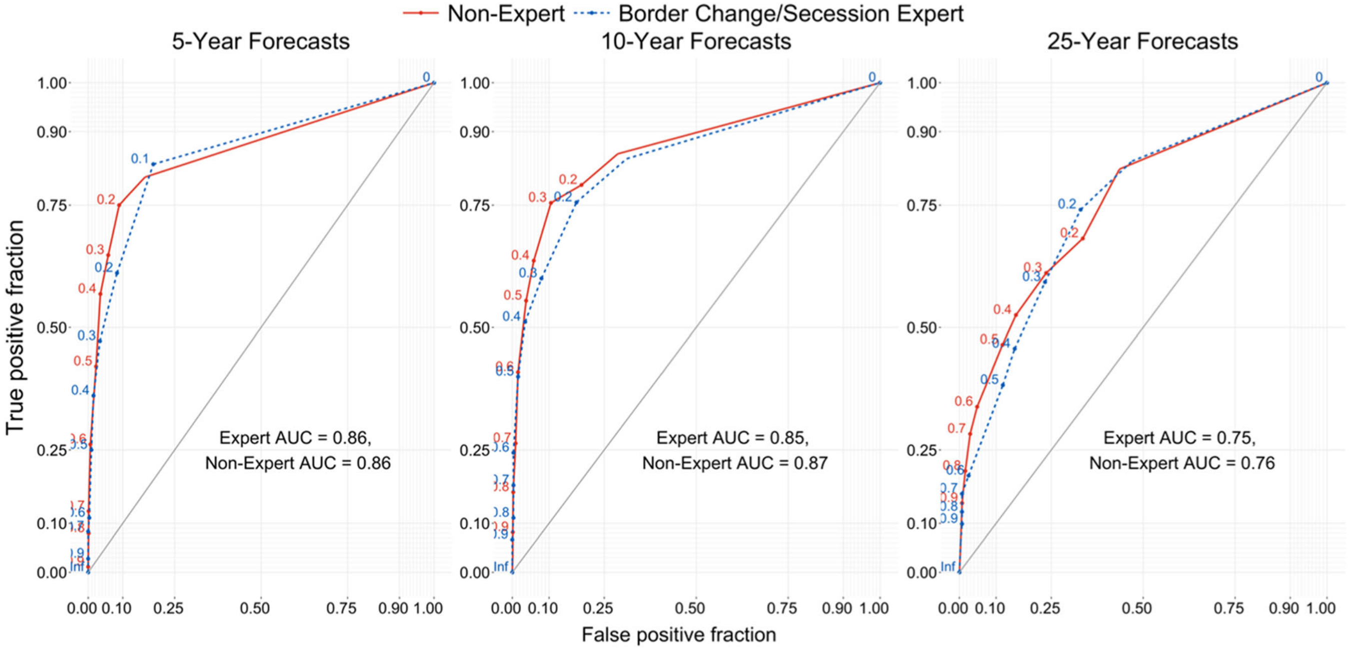

FIGURE 2 ROC analysis of border change/secession forecasts of experts and non-experts across 5, 10, and 25 years. AUC, area under the curve; ROC, receiver operating characteristic.

proliferation forecast periods but not for the border change domain, where the red and blue curves intersect in at least two plots. The dominance of the sort achieved by proliferation experts is an impressive achievement. It implies that false positives accrue slower for them at all points across the entire probability scale—and that decision-makers should always prefer experts' forecasts, regardless of where they choose to set their threshold for acting on forecasts.

A second performance metric is the area under the curve (AUC) integral. AUC scores appear in each panel of the two Figures and can range from perfect discrimination (AUC = 1) to pure-guessing accuracy (AUC = 0.5), represented by the straight diagonal. AUC scores let us compare groups within and across domains. Of course, how impressed we are by a given AUC score hinges on our assumptions about task difficulty and skill levels of forecasters.

Here we make three additional points. First, notwithstanding media chatter about how the average expert is no better than a dart-tossing chimp, individual forecasters outperformed chance across the board (Tetlock, 2017). Experts and non-experts alike bested the chimps (AUC = 0.5) by wide margins for nuclear proliferation (all AUC > = 0.89) and even border change (all AUC > = 0.75). The magnitude of these effects suggests either: (a) most questions were easy relative to those posed about faster-moving variables in other tournaments (Satopää et al., 2021); (b) the professionals in these studies were unusually gifted forecasters. Our best bet is that it is due to a mix of these two factors.

Second, the AUC for border change was smaller than for nuclear proliferation—a sign that experts and generalists alike found border change a harder task. Indeed, the toughest proliferation task—25-year forecasts—yielded AUC values of 0.97 for specialists and 0.89 for generalists, whereas the easiest border-change task—5-year forecasts—yielded AUC scores of 0.86.

Third, experts' edge over generalists in the proliferation domain grew the more distant the predictive horizon. Long-range forecasting did not get much harder for proliferation experts; their AUC scores declined only a little: from 0.98 to 0.97. But the generalists dropped from 0.93 to 0.89.

These signal-detection results imply greater causal regularity in the proliferation domain, but regularities not so glaring that generalists could pick them up as readily as specialists. In the messier border-change domain, specialists' and generalists' accuracy fell off more steeply: from 0.86 to 0.75.

An open question for a follow-up tournament for 2024–2049 is whether we can experimentally create instant experts in high-causal-regularity domains by adapting checklist methods that itemize well-defined steps to success—a method that has worked in fields like surgery and aviation (Gawande, 2010). Applied to nuclear proliferation, key questions might be: Which suspect states possess the technological-economic means? And do leaders of those states see the domestic and geopolitical incentives to “go nuclear” as outweighing the disincentives of international sanctions? Checklist manipulations should prove less helpful in low-causal-regularity domains like border change, where there are many messy pathways to the outcomes and expertise is diffused across too many idiosyncratic situations. How many specialists claim expertise on secessionist politics across Sri Lanka, Nigeria, Iraq, Canada, Spain,… and border disputes between Congo-Rwanda, Russia-Ukraine, China-India,…?

The replicability of 1988/97–2023 results in 2024–2049 will hinge on both methodological details under researchers' control and historical constraints outside their control. If researchers pose too easy questions, virtually everyone gets everything right, and AUCs approach 1.0. It is safe to predict that Norway will not embrace nuclear weapons. Or if they pose questions that are too difficult—if all proliferation questions had been as hard as Iran, a close-call state—the value of their expertise would fall close to zero. Question drafters need to aim for the Goldilocks zone of difficulty, which turns them into forecasters themselves.

Our expectations for replicability will also hinge on our working philosophy of history. Scholars range in their tolerance for counterfactual scenarios, from determinists who reject them to probabilists who embrace them, but the stubborn fact is that no one knows how close we came to an alternative world in which, say, Iran went nuclear (which, ex-ante, the best forecasters thought quite likely). Recoding Iran as a 1.0 (proliferator) rather than a zero (nonproliferator) would not by itself erase experts' edge but if a nuclear-armed Iran triggered a chain reaction in the Arab world and beyond, experts' edge would vanish fast. How impressed we are by objective forecasting accuracy rests on subjective counterfactual assumptions (Tetlock & Belkin, 1996).

Here Solingen's (2023) commentary offers valuable ideas for follow-up work. How exactly did specialists manage to maintain both a high hit rate and a low false-positive rate? Did their advantage lie more in spotting true-positives (e.g., North Korea) or in resisting false-positive judgments? Solingen proposes a testable hypothesis as to why generalists may have a harder time discounting superficially plausible proliferation suspects.

Solingen conjectures that generalists took an accuracy hit because they were thinking, as were most scholars in world politics at the time, like top-down neo-realists and were treating forecasting problems as covering-law syllogisms. The major premise would be that world politics is anarchic and states require the means to defend themselves on their own or with the help of powerful protectors. The minor premise would be that, over time, states worry about the reliability of their protectors and opt to acquire their own nuclear deterrent capacity. The conclusion is that we should expect nuclear proliferation to extend to a far wider range of nation-states than the nine thus far in 2023—perhaps closer to John F. Kennedy's projection in 1963: 25 or more.

If generalists in our sample had been thinking along these neorealist lines, then—relative to specialists—they should have inflated their proliferation probabilities for states that had hostile nuclear-armed neighbors and that possessed the economic-technological capacity to construct nuclear weapons: Germany, Japan, South Korea, and Taiwan—plus the Middle Eastern suspects of Iran, Egypt, Iraq, and Saudi Arabia. Table 1 shows the average 25-year proliferation forecasts collapsed across the 1988 and 1997 elicitation periods. Although generalists did not make big false-positive bets on these eight states, they made higher bets than specialists did (with the close-margin exceptions of Germany and Iran). On the flipside, generalists made weaker bets on the three true-positive proliferators than did specialists, which supports Solingen's view that good judgment in the proliferation domain required a multidimensional worldview that appreciated not just the neorealist self-help pressures to go nuclear but also the cultural and domestic political pressures to inhibit nuclear-weapons development as well as the power of international sanctions (and military force) to thwart weapons programs. In brief, better proliferation forecasters were likelier to be eclectic foxes than neo-realist hedgehogs.

Finally, Poast (2023) raises a foundational challenge to the narrow accuracy-quantification agenda of psychometric studies of

TABLE 1 Average 25-year forecasts for nuclear proliferation.

Location	Proliferation indicator	Expert avg. forecast (SD)	Nonexpert avg. forecast (SD)
Taiwan*	0	0.01 (0.04)	0.07 (0.13)
Egypt*	0	0.04 (0.06)	0.08 (0.12)
Japan*	0	0.05 (0.09)	0.10 (0.13)
South Korea*	0	0.05 (0.07)	0.10 (0.12)
Germany*	0	0.06 (0.15)	0.05 (0.11)
Saudi Arabia*	0	0.06 (0.09)	0.14 (0.16)
Iraq*	0	0.37 (0.25)	0.49 (0.24)
Iran*	0	0.52 (0.25)	0.46 (0.24)
Pakistan	1	0.71 (0.27)	0.51 (0.27)
North Korea	1	0.72 (0.25)	0.52 (0.24)
India	1	0.75 (0.24)	0.54 (0.14)
Other 12 Locations	0	0.04 (0.09)	0.08 (0.15)

Note: Asterisks denote non-nuclear weapons states that, from a neorealist view, constitute long-run risks.

Abbreviation: SD, standard deviation.

forecasting. Policy-makers might be interested that the current crowd probability of Iran going officially nuclear in 2024 is, say, 0.3 but they would find it vastly more useful to know the factors that could either reduce or inflate that risk. How much would the odds change if the USA reduced the sanctions Iran most resents (doves' forecast) or if the United States intensified sanctions and issued a cease-and-desist ultimatum backed by the threat of a massive pre-emptive attack (hawks' forecast)?

The list of conditionals is endless. And Poast notes that the big questions tend to be embedded in intertwined webs of conditionals. Looking back over the last century for far-sighted or far-off forecasts, assessments of their accuracy pivot on often hidden assumptions. Keynes' (1919) "The Economic Consequences of the Peace" is widely upheld as amazingly prescient in anticipating the rise of German revanchism and World War II, but the connections between these abstract risk factors and specific historical events—like the rise of the Nazi Party and the adoption of appeasement policies that facilitated the consolidation of German strength in the 1930s—are highly contingent and beyond even the shrewdest observers' capacity to anticipate in 1919. Many critics of the Treaty of Versailles may well have later been ardent advocates of appeasing Hitler.

In sum, our commentators convince us there is a wide rigor-relevance gap between what specialists in subjective-probability forecasting have to offer and specialists in world politics need. Narrowing this gap has proven hard. We work at such different levels of analysis. But we should try. The net result would be a research literature more relevant to decision-makers but still rigorous in the eyes of the research community. There is plenty of room for Pareto improvement.

ACKNOWLEDGMENTS

We thank two foundations—Open Philanthropy and Founders Pledge—for their generous research support.

DATA AVAILABILITY STATEMENT

Data availability upon request.

ORCID

Philip E. Tetlock http://orcid.org/0000-0002-3199-0292
Kevin Chen http://orcid.org/0009-0007-8307-5962

REFERENCES

Gawande, A. (2010). *The checklist manifesto: How to get things right*. Metropolitan Books.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

Poast, P. (2023). Prediction in international relations is hard, sometimes. *Futures & Foresight Science*. Advance online publication.

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. https://doi.org/10.1287/mnsc.2020.3882

Solingen, E. (2023). Nuclear cascades or more of the same? Why meliorists may have gotten it right: A commentary on Tetlock at al. (2023). *Futures & Foresight Science*, e168. Advance online publication. https://doi.org/10.1002/ffo2.168

Tetlock, P. E. (2017). *Expert political judgment: How good is it? How can we know?* (2nd ed.). Princeton University Press.

Tetlock, P. E. & Belkin, A. (Eds.). (1996). *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press.

Tetlock, P. E., Karvetski, C., Satopää, V. A., & Chen, K. (2023). Long-range subjective-probability forecasts of slow-motion variables in world politics: Exploring limits on expert judgment. *Futures & Foresight Science*, e157. Advance online publication. https://doi.org/10.1002/ffo2.157

Treisman, D. (2023). What is predictable? A commentary on Tetlock at al. (2023). *Futures & Foresight Science*, e166. Advance online publication. https://doi.org/10.1002/ffo2.166

How to cite this article: Tetlock, P. E., Karvetski, C., Satopää, V. A., & Chen, K. (2024). Exploring the limits on Meliorism: A commentary on Tetlock et al. (2023). *Futures & Foresight Science*, *6*, e173. https://doi.org/10.1002/ffo2.173