COMMENTARY

Judging political judgment

Philip Tetlock[1] and Barbara Mellers

Department of Psychology and The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

Mandel and Barnes (1) have advanced our understanding of the accuracy of the analytic judgments that inform high-stakes national-security decisions. The authors conclude that, in contrast to past work (2), the experts they studied (Canadian intelligence analysts) make surprisingly well-calibrated, high-resolution forecasts. We worry, however, about apple-orange comparisons.

Multidimensional Comparisons

The relatively poor performance in Tetlock's earlier work was most pronounced for long-term forecasts (often 5 y plus) and among forecasters who had strong theoretical priors and did not feel accountable for their judgments. These are favorable conditions for generating overconfidence. In contrast, Mandel and Barnes (1) found favorable conditions for generating well-calibrated and high-resolution probabilistic judgments. The authors studied much shorter-term forecasts (59% under 6 mo and 96% under a year), and their forecasters worked not under the anonymity guarantees given human subjects but rather under accountability pressures designed to enhance judgment (3, 4).

Suggestive support for this analysis emerges from a massive geopolitical forecasting tournament sponsored by the Intelligence Advanced Research Projects Activity (IARPA). Our research group (5, 6) won this tournament and found, using time frames similar to those in Mandel and Barnes (1), that its best forecasting teams achieved Brier scores similar to those of Canadian analysts. The tournament also permits randomized experiments that shed light on how to design conditions—training, teaming, and accountability systems—for boosting accuracy (5). These efforts implement a key recommendation of a 2010 National Academy Report: start testing the efficacy of the analytical methods that the government routinely purchases but rarely tests (7, 8). According to David Ignatius of the Washington Post, these efforts have already produced a notable upset: the best practices culled from the $5 million-per-year IARPA tournament have generated forecasts that are reportedly more accurate than those generated by the intelligence community (9), whose total annual funding is well in excess of $5 billion.

Acknowledging Our Ignorance

We should, however, focus on the core problem that neither past nor current work has yet solved: how best to measure the deceptively simple concept of accuracy. One challenge is the standardization of difficulty. Getting a good Brier score by predicting weather in a low-variance world (e.g., Phoenix) is a lot easier than it is in a high-variance world (e.g., St. Louis) (10). When forecasters across studies answer questions of varying difficulty embedded in historical periods of varying predictability, cross-study comparisons become deeply problematic.

Mandel and Barnes have advanced our understanding of the accuracy of the analytic judgments that inform high-stakes national-security decisions.

Mandel and Barnes (1) focused on questions that analysts could answer almost perfectly, yielding Brier scores of 0 or 0.01 over half of the time, which requires assigning 0s and 0.1s to nonoccurrences and 1s and 0.9s to occurrences. Their subject-matter experts rated the difficulty of questions retrospectively and classified 55% of questions as "harder." However, this begs the question: Harder than what?

In our view, ratings of question difficulty are best done ex ante to avoid hindsight bias, and this rating task is itself very difficult because we are asking raters, in effect, to predict unpredictability (11, 12). The forecasts labeled "hard" in Mandel and Barnes (1) may be quite easy [relative to Tetlock (2)], and the forecasts they label "easy" may be very easy [relative to Mellers et al. (5)], or we may not know the true difficulty for decades, if ever. Suppose a rater classifies as "easy" a question on whether there will be a fatal Sino-Japanese clash in the East China Sea by date *X*, and the outcome is "no." Should policy-makers be reassured? Two major powers are still playing what looks like a game of Chicken, which puts us just one trigger-happy junior-officer away from the question turning into a horrendously hard one. "Inaccurate" forecasters who assigned higher probabilities may well be right to invoke the close-call counterfactual defense (it almost happened) and off-on-timing defense (wait a bit longer. . .) (2).

Another problem, which also applies both to our work and to Mandel and Barnes (1), is that Brier scoring treats errors of under- and overprediction as equally bad (13). However, that is not how the blame game works in the real world: underpredicting a big event is usually worse than overpredicting it. The most accurate analysts in forecasting tournaments—those who were only wrong once and missed World War III—should not expect public acclaim.

Reducing Our Ignorance

Mandel and Barnes are right. Tetlock (2) did not establish that analysts are incorrigibly miscalibrated, and we would add that Mandel and Barnes (1) and Mellers et al. (5) have not shown they are typically well calibrated. We need to sample a far wider range of forecasters, organizations, questions, and time frames. Indeed, we do not yet know how to parameterize these sampling universes. All we have are crude comparisons (group *A* working under conditions *B* making forecasts in domain *C* in historical period *D* did better than . . .).

Intelligence agencies rarely know how close they are to their optimal forecasting frontiers, along which it becomes impossible to achieve more hits without incurring false alarms. When intelligence analysts are forced by their political overseers

Author contributions: P.T. and B.M. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 10984 of issue 30 in volume 111.

[1]To whom correspondence should be addressed. Email: tetlock@wharton.upenn.edu.

into spasmodic reactions to high-profile mistakes—by critiques such as, "How could you idiots have missed this or false alarmed on that?"—the easiest coping response is the crudest form of organizational learning: "Whatever you do next time, don't make the last mistake." In signal detection terms, you just shift your response threshold for crying wolf (4).

Keeping score and testing methods of boosting accuracy facilitates higher-order forms of learning that push out performance frontiers, not just shift response thresholds. Although interpreting the scorecards is problematic, these problems are well worth tackling, given the multitrillion-dollar decisions informed by intelligence analysis.

ACKNOWLEDGMENTS. This research was supported by the Intelligence Advanced Research Projects Activity via the Department of Interior National Business Center Contract D11PC20061.

1 Mandel DR, Barnes A (2014) Accuracy of forecasts in strategic intelligence. *Proc Natl Acad Sci USA* 111(30): 10984–10989.
2 Tetlock PE (2005) *Expert Political Judgment: How Good Is It? How Can We Know?* (Princeton Univ Press, Philadelphia, PA), 321 pp.
3 Lerner JS, Tetlock PE (1999) Accounting for the effects of accountability. *Psychol Bull* 125(2):255–275.
4 Tetlock PE, Mellers BA (2011) Intelligent management of intelligence agencies: Beyond accountability ping-pong. *Am Psychol* 66(6):542–554.
5 Mellers B, et al. (2014) Psychological strategies for winning a geopolitical forecasting tournament. *Psychol Sci* 25(5):1106–1115.
6 Tetlock PE, Mellers B, Rohrbaugh N, Chen E (2014) Forecasting tournaments: Tools for increasing transparency and the quality of debate. *Curr Dir Psychol Sci*, 10.1177/0963721414534257.
7 National Research Council (2011) *Intelligence Analysis for Tomorrow: Advances from the Behavioral and Social Sciences* (National Academies Press, Washington, DC), 102 pp.
8 Fischhoff B, Chauvin C, eds (2011) *Intelligence Analysis: Behavioral and Social Scientific Foundations* (National Academies Press, Washington, DC), 338 pp.
9 Ignatius D (Nov 1, 2013) More chatter than needed. *The Washington Post*. Available at http://www.washingtonpost.com/opinions/david-ignatius-more-chatter-than-needed/2013/11/01/1194a984-425a-11e3-a624-41d661b0bb78_story.html.
10 Murphy AH, Winkler RL (1987) A general framework for forecast verification. *Mon Weather Rev* 115(7):1330–1338.
11 Fischhoff B (1975) Hindsight not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *J Exp Psychol Hum Percept Perform* 1(5):288–299.
12 Jervis R (2010) *Why Intelligence Fails: Lessons from the Iranian Revolution and the Iraq War* (Cornell Univ Press, Ithaca, NY).
13 Green DM, Swets JA (1966) *Signal Detection Theory and Psychophysics* (Wiley and Sons, New York, NY).