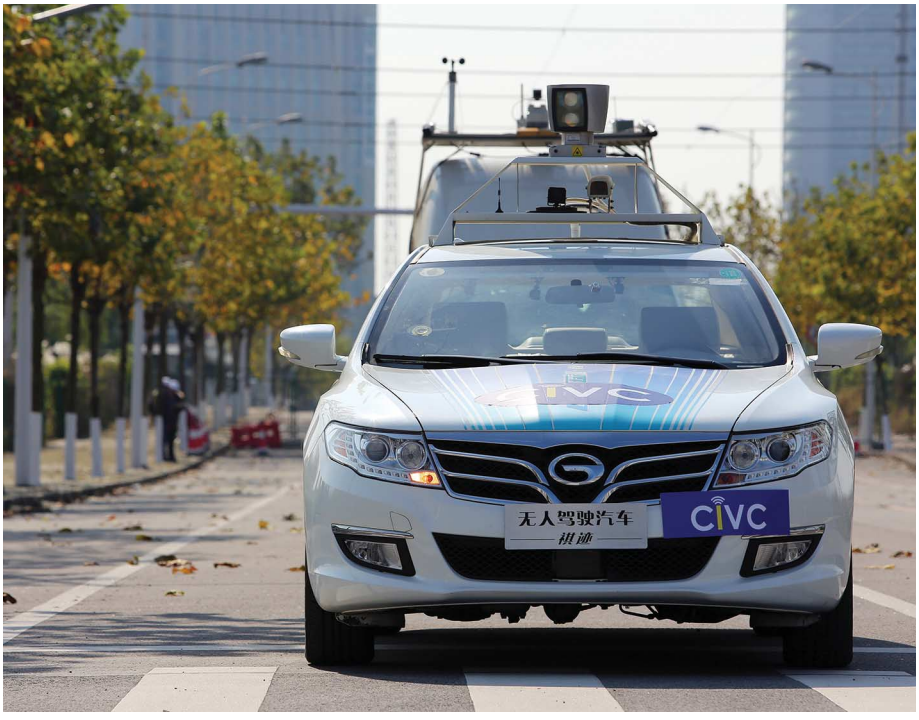

PHOTO: IMAGINECHINA VIA AP IMAGES

What is the future for driverless cars? A driverless car travels on the road during the 2016 China Intelligent Vehicle Championship in Shanghai, China, 3 December 2016.

ESSAY

Bringing probability judgments into policy debates via forecasting tournaments

Philip E. Tetlock,[1]* Barbara A. Mellers,[1] J. Peter Scoblic[2]

Political debates often suffer from vague-verbiage predictions that make it difficult to assess accuracy and improve policy. A tournament sponsored by the U.S. intelligence community revealed ways in which forecasters can better use probability estimates to make predictions—even for seemingly "unique" events—and showed that tournaments are a useful tool for generating knowledge. Drawing on the literature about the effects of accountability, the authors suggest that tournaments may hold even greater potential as tools for depolarizing political debates and resolving policy disputes.

Whether the topic is national security, interest rates, or environmental legislation, policy debates often hinge on competing claims about the probability of predicted consequences. However, holding partisans accountable for their forecasts is virtually impossible because they frame them in vague verbiage: Raising interest rates "may" trigger a recession; providing military assistance to Ukraine "could" provoke a sharp Russian response; or environmental legislation "might" increase energy prices. When readers are asked to translate the words in quotation marks into probability judgments, the answers straddle both sides of "maybe," taking on meanings as low as a 0.08 chance of occurrence to some, and as high as 0.59 to others (*1*). If the outcome does occur, the prognosticator can say, "I warned you that it was a distinct possibility," and if it does not, he can shrug and say, "I merely said it was possible." Vagueness thus precludes accountability, which in turn impedes our ability to learn and to improve the accuracy of our forecasts. And if we cannot improve our forecasts, we make it that much more difficult to improve policy.

In contrast, forecasting tournaments—contests among individuals or teams—address each of these problems by incentivizing competitors to make accurate predictions about specific events. In 2011, the U.S. Intelligence Advanced Research Projects Activity's (IARPA's) Aggregative Contingent Estimation (ACE) program held a 4-year forecasting tournament to experiment with generating numerical probability estimates. The U.S. intelligence community has long relied on vague-verbiage forecasting (*2*), but in the IARPA tournament, five university-based teams that responded to the agency's request for contestants competed to produce the most accurate predictions on a wide array of geopolitical and economic topics, ranging from the performance of financial markets, to the risk of Greece leaving the Eurozone, to the prospects of a violent Sino-Japanese clash in the East China Sea (*3*). IARPA formulated hundreds of questions about such topics so that they were "resolvable"—they could be answered "yes" or "no" within a specified time frame—and scored performance using Brier points, which are calculated as the sum of the squared errors between a probability forecast and reality (which can be coded "1" for events that happened and "0" for events that did not). So, for example, a forecaster might have predicted that there was an 80% chance that the Dow Jones Industrial Average would finish the year above 20,000 (and, by extension, a 20% chance that it would not). Because the Dow did not pass 20,000, the forecaster's Brier score would be calculated as $(0.8 - 0)^2 + (0.2 - 1)^2 = 1.28$. This approach yields scores between 0 (perfect omniscience) and 2 (total failure), and because it squares errors, a Brier score rewards decisive correctness while steeply punishing overconfidence. The specificity of the tournament's questions enabled accountability, the feedback provided by Brier scores enabled learning, and learning ultimately improved accuracy.

In conducting its research, one of the teams competing in the IARPA tournament, the Good Judgment Project (GJP; co-created by authors P.E.T. and B.A.M.), held competitions among its experimental subjects—some 2400 Americans with a wide range of demographic and professional backgrounds—designed to elicit their most accurate probability forecasts. In other words, it held tournaments within the overall ACE tournament that it then used to produce team predictions. There were several key findings.

1) Some methods for extracting wisdom from crowds are better than others. Prediction polls yield a probabilistic forecast by aggregating the predictions of individuals through a range of methods, from simple averaging (which outperforms most individual forecasts) to fancier tools, such as log-odds extremizing of weighted averages (which works even better). In contrast, prediction markets rely on forecasters buying and selling contracts whose ultimate value depends on the outcome of a future event. For example, a bettor who believes that Candidate X is at least 80% likely to win his election might purchase a $1 futures contract for 80 cents from a seller who believes that the probability of victory is 20% or less. If the candidate wins, the seller owes the buyer 20 cents, and if the

[1]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA. [2]Harvard Business School, Harvard University, Boston, MA 02163, USA.
*Corresponding author. Email: tetlock@wharton.upenn.edu

candidate loses, the buyer must pay the seller 80 cents. As contract prices fluctuate in response to supply and demand, so does the implied probability of the event in question. A 3-year series of randomized control trials showed that such markets outperformed the simple crowd average of prediction polls, but more complex weighted-averaging methods of distilling crowd wisdom outperformed markets (*4*).

2) The winning algorithm across all tournament years was a log-odds weighted-averaging equation that extremized median probability judgments (for example, transforming 0.7 into 0.85 or 0.3 into 0.15) as a function of the diversity of the views feeding into the median. The rationale is intuitive. Imagine the director of the Central Intelligence Agency asks her advisers for independent estimates of the likelihood that a terrorist mastermind is in a certain location and that each adviser gives the same answer: $P = 0.7$. What should the director conclude? It depends. If the advisers are clones of each other, drawing on the same evidence and analyzing it through the same opinion prism, the answer is indeed 0.7. But if different advisers draw on different evidence—one on satellite imagery, another on cyber-intelligence, and the third on human informants—then the director now has grounds for extremizing beyond 0.7. How much to extremize can be statistically estimated if you have a rich database on the predictive track records of your advisers.

3) Some forecasters are, surprisingly consistently, better than others, and we now know a lot about the top performers. Their personality profiles revealed above-normal scores on measures of active open-mindedness (a willingness to treat one's beliefs as testable hypotheses, not sacred possessions), on measures of cognitive growth mindset (a willingness to treat forecasting as a skill that can be cultivated and is worth cultivating), and on measures of acceptance that chance plays a key role in shaping life outcomes (a skepticism of efforts to imbue life-altering coincidences with deep meaning, such as fate) (*5*). The most successful forecasters did tend to have advanced degrees and a greater degree of general political knowledge, but subject-matter expertise itself conferred little benefit because the questions asked during the tournament covered such a wide range of topics.

4) Learning—and therefore improvement—is possible, even though the world of international politics and economics, in which IARPA was interested, is not learning-friendly. Unlike poker, which involves random draws from a well-defined sampling universe, with rapid feedback on accuracy, global events are often one of a kind. We cannot test the precision of probabilities by rerunning history from a designated start point and observing, for example, how often Russia annexes Crimea or Greece leaves the Eurozone. Some scholars even deny that probabilistic forecasts of unique situations can have any meaning (*6*). Our view is evidence-based: The ACE tournament demonstrated forecasters' capacity to learn to make well-calibrated probability judgments about just such situations. In fact, top forecasters strove to make their probability judgments as granular as possible, and other studies have shown the utility of quantifying probabilities, even of global events (*7*, *8*). Misunderstandings about probabilistic forecasts underplay this achievement. Many people do not appreciate that when a perfectly calibrated system offers a prediction of 70%, then 30% of the time the system will be "wrong."

Leveraging these findings allowed GJP to generate forecasts that outperformed—by roughly 30%—a prediction market run by the U.S. intelligence community in which the players were professional analysts with access to classified information (*3*–*5*, *9*–*11*). By producing a superior forecasting methodology, the ACE tournament yielded an important public policy tool: If policy-makers have access to more accurate forecasts, they can better anticipate the consequences of their actions and therefore make better decisions.

"Forecasting tournaments—contests among individuals or teams—[incentivize] competitors to make accurate predictions about specific events."

More generally, the IARPA contest demonstrated the utility of tournaments as a tool for knowledge production. GJP's tournaments within the ACE competition allowed randomized-control trials of how best to boost accuracy. These experiments demonstrated the surprising effectiveness of short training or debiasing exercises that taught forecasters how to ground probability estimates in base rates and to update their beliefs in a roughly Bayesian fashion in response to new evidence. Other experiments demonstrated the power of well-choreographed forms of teamwork. Training team members how to precisely but diplomatically question each other's assumptions—how to disagree without being disagreeable—helped groups outperform the same number of individuals working alone. Tournaments thus proved themselves a useful method for conducting experiments outside the laboratory.

We suspect that tournaments can do even more by providing a framework for resolving public policy debates. A key feature of tournaments is accountability—participants in the GJP tournaments were publicly ranked according to the accuracy of their forecasts—and research has shown that predecisional accountability prompts individuals to engage in preemptive self-criticism (*12*, *13*). Faced with the prospect of having to justify a position or decision, they consider the ways in which their audience might react. This effort increases cognitive complexity, by which individuals contemplate a greater number of germane factors—or, in the case of a political problem, arguments for or against a particular position. Having considered a wider range of views and anticipating a critical audience, individuals may moderate their beliefs. Were political opponents to participate in a forecasting tournament, they might well temper their predictions and, by implication, the extremeness of their policy positions.

Admittedly, the specificity required to make questions rigorously resolvable precludes asking "big" questions. Take a concern such as "Will increasing automation destabilize white-collar labor markets over the next 20 years?" Here, key terms are too ill-defined for the question to ever be answered "yes" or "no." Probability estimates are therefore unhelpful. But there is a way around this problem: Subject-matter experts can generate specific resolvable indicators, each of which captures a distinct facet of the larger issue. For instance, will driverless for-hire vehicles without human supervision be taking passengers in a major American city by the end of 2017? Will IBM's Watson engine outperform top physicians in medical-diagnosis tournaments by the end of 2018? Will half of all accounting jobs be automated by the end of 2019? Each micro question has independent diagnostic value vis-à-vis the macro question. Taken together, these "Bayesian question clusters" can help policy-makers better prepare for looming megachanges.

Imagine, then, forecasting tournaments in which political opponents compete to answer Bayesian question clusters. The accountability inherent in the tournament framework could promote cognitive complexity, while the question clustering would enable participants to tackle fundamental differences. Forecasting tournaments could therefore depolarize political debate—even about major issues. Here, we speculate, but proof of concept can be found in the literature.

Building on Kahneman's concept of adversarial collaboration—in which opposing camps agree to empirically test competing claims, clarifying ahead of time what evidence would challenge their existing conceptions and therefore change their minds (*14*, *15*)—clashing camps could nominate question clusters that they believe they are better equipped to answer. For instance, in a tournament designed to gauge the value of the 2015 deal limiting Iran's nuclear program, hawks might nominate questions on the continued dominance of hardliners in government, whereas doves might focus their questions on the benefits of opening Iran's economy. Inaccurate predictions about one's own questions could force reconsideration of one's policy position. The competitors could strengthen the effectiveness of this exercise if they agreed to make predictions about the same germane questions—say, Iran's compliance with

International Atomic Energy Agency inspections. By agreeing on the indicators that have the most diagnostic value, participants would establish predecisional accountability that would be difficult to evade if their predictions broke the wrong way.

The focus of GJP was on getting as much out of humans as possible. The next generation of planned tournaments will explore human-machine hybrids—and the types of questions on which people can add value beyond machines. The working hypothesis is that machines become ever harder to beat when tournaments pose questions about criterion variables with long quantifiable time series and extensive correlations with networks of other quantifiable variables (for example, "What will gross domestic product growth be in the UK in the first quarter of 2017?"). But machines stumble when we pose questions that are high on the uniqueness continuum (that lack easy-to-specify comparison classes that algorithms can use to generate probability estimates), such as, "What is the likelihood of a general election being called in the UK before 31 March 2017?" The latter question is higher on the uniqueness scale because historical precedents are much harder to quantify and define, and the decision ultimately pivots on a complex case-specific decision calculus.

No single exercise will resolve a high-profile debate. But as long as one can get away with making vague-verbiage predictions, one never has to admit that one was wrong. In contrast, if events to which one assigns 80% probabilities keep occurring 20% of the time, one starts to lose wiggle room, and it becomes much easier for the rest of us to figure out which points of view to assign greater credibility in which policy domains. Tournaments could, in this way, help to pry open otherwise closed minds—and depolarize unnecessarily polarized debates.

REFERENCES AND NOTES

1. F. Mosteller, C. Youtz, *Stat. Sci.* **5**, 2–12 (1990).
2. S. Kent, *Studies Intell.* **8**, 49–65 (1964).
3. B. Mellers *et al.*, *Psychol. Sci.* **25**, 1106–1115 (2014).
4. P. Atanasov *et al.*, *Manage. Sci.* 10.1287/mnsc.2015.2374 (2016).
5. B. Mellers *et al.*, *J. Exp. Psychol. Appl.* **21**, 1–14 (2015).
6. N. Taleb, M. Blyth, *Foreign Aff.* **90**, 33–39 (2011).
7. A. Barnes, *Intell. Natl. Secur.* **31**, 327–344 (2015).
8. T. S. Wallsten, D. Budescu, A. Rapoport, R. Zwick, B. Forsyth, *J. Exp. Psychol. Gen.* **115**, 348–365 (1986).
9. B. Mellers *et al.*, *Perspect. Psychol. Sci.* **10**, 267–281 (2015).
10. P. Tetlock, D. Gardner, *Superforecasting: The Art and Science of Prediction* (Crown, 2015).
11. P. Tetlock, B. Mellers, N. Rohrbaugh, E. Chen, *Curr. Dir. Psychol. Sci.* **23**, 290–295 (2014).
12. P. Tetlock, *J. Pers. Soc. Psychol.* **45**, 74–83 (1983).
13. P. E. Tetlock, L. Skitka, R. Boettger, *J. Pers. Soc. Psychol.* **57**, 632–640 (1989).
14. B. Mellers, R. Hertwig, D. Kahneman, *Psychol. Sci.* **12**, 269–275 (2001).
15. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).

ACKNOWLEDGMENTS

The authors thank the Carnegie Corporation and Open Philanthropy Network for funding.

10.1126/science.aal3147

Bringing probability judgments into policy debates via forecasting tournaments

Philip E. Tetlock, Barbara A. Mellers and J. Peter Scoblic

Science **355** (6324), 481-483.
DOI: 10.1126/science.aal3147