©ISTOCKPHOTO.COM/BERT HEYDEL

Crowd wisdom

U.S. intelligence researchers seek public help with forecasting experiment

By Michael Peck

Anyone can make a prediction. But not everyone can do it well, a fact of life that has had dismal implications for forecasting by intelligence agencies.

That's why American researchers have begun a project to identify those intelligence analysts who should be listened to more than others when policymakers ask for predictions about the outcome of an event. The project, called ACE for Aggregative Contingent Estimation, is sponsored by the Intelligence Advanced Research Projects Activity (IARPA), an agency set up after the Sept. 11 terrorist attacks to spur innovation. For one, IARPA wants to see if it is possible to assign more weight to the predictions of successful forecasters to improve aggregate forecasts.

"It has long been known that simply averaging a large set of independent judgments creates an estimate that is usually more accurate than most of the individual judgments and sometimes more accurate than every individual judgment in the group," said IARPA's Jason Matheny, program manager for ACE. Intelligence officials call the phenomenon the wisdom of crowds.

IARPA has hired five teams to explore how the phenomenon might be applied to improving the community's ability to forecast outcomes. If things go well in the first year, IARPA could extend the research across four years.

Some of the researchers are asking members of the public to participate. Researchers want to see if breaking questions about large events into questions about smaller events improves estimates. Others will test the weighting concept by tracking the performance of participants and assigning more value to the views of successful estimators. Each team will be given around 100 problems per year, and they'll be scored on accuracy against real-world events.

The teams are led by Applied Research Associates, Draper Laboratory, Jacobs Technology, Virginia's George Mason University, the University of California, Berkeley, and the University of Pennsylvania.

A danger of lumping multiple forecasts into an aggregate forecast is that junk predictions are weighed equally with good ones. IARPA wants to see if it is possible to identify the better forecasters and use their estimates to improve the aggregate prediction. Say, for example, that there's an estimate being prepared for Congress on when Iran will develop a nuclear bomb. If 10 analysts estimate that it will be 2015 and five say that it will be 2020, but those five analysts have qualities associated with successful forecasting, then their estimates would be weighed equal to the views of the 10 analysts.

"If we are able to pick up decisions about how people do forecasting, and then find the telltale signs that someone is engaging in those types of activities, we might put more weight on those people," said Dirk Warnaar, principal investigator for ACE research at Applied Research Associates in Raleigh, N.C.

SOCIAL MEDIA

To find the qualities of good forecasters, the ARA team, which includes researchers from seven universities, is using Twitter, Facebook and the Web to go outside the professional forecasting world. In June, the researchers set up a website — http://forecastingace.com —

inviting volunteers to register to make predictions about the outcomes of events lasting weeks or months. For one group, researchers will gradually place more weight on predictions from those who are proved right. The accuracy of this group's aggregate predictions will be compared over time with those from an unweighted group. ARA is hoping for 1,000 to 10,000 volunteers for the crowdsourcing experiment, which could last one to four years. IARPA is looking for a 20 percent improvement in forecasting accuracy for the weighted group compared with the unweighted group in the first year of the project, leading up to a 50 percent increase in the fourth year, should the project last that long.

ARA's site asks volunteers to pick questions from military, political, scientific, social and economic topics. Under the military topic is the question, "Will the number of U.S. troops in Afghanistan reduce in July 2011?" If the participant answers "yes," then the site asks the participant for an estimate of confidence on a scale of 50 to 100 percent, and for details of the considerations that went into the prediction.

Another question might be, "Will Iran test a nuclear weapon by 2013?" Users would be asked if they could think of any factors related to the questions, such as Iran's capabilities, intentions, the steps involved in testing a weapon, and the political ramifications of a test. This input could be used to generate subquestions such as, "Does Iran have refined weapons-grade plutonium? Does it have a suitable test site? Can each step be completed by 2013?"

"By asking the subquestions, we can now determine on which aspects there is agreement and disagreement among the participants," Warnaar said. "Besides the overall probability for the target question, the level of agreement and disagreement will be very useful for the end-user of the forecast, a decision-maker such as a member of Congress."

ARA is billing ACE as a win-win project; the government gets better forecasting, and volunteers get self-improvement.

"In applying these methods, we should be able to tell which participants are better forecasters and why," Warnaar said. "We will share that information with each participant, which they could apply on subsequent forecast problems. We believe that a participant's self-awareness will improve their forecasting performance. Awareness and self-improvement could also have significant benefits for participants in their daily lives. For example, if it turns out that you are generally overconfident, you may take this into account when you make your next investment decision or when you estimate how much time it takes to drive across town."

Jennifer Carter, who directs ARA's public relations for the project, doesn't expect that finding volunteers will be a problem.

"There are so many communities out there already interested in forecasting that this will provide a medium for them to explore that interest. I think, through Facebook and Twitter, we are going to find people who are into social media, who are readers of blogs and online media, and these people tend to be up-to-date on current events," she said. "And while they may not have a particular focus on politics or economics or technology, they know a little bit about everything. Because they're up-to-date and they're always online, they'll probably make really good forecasters."

QUESTIONS AND METHODOLOGIES

Yet the underlying concept of ACE raises a host of questions. Which variables should be included when assessing the influences on the accuracy of someone's forecast? Is it college versus graduate school education? Do engineers make better predictions than political science majors? Warnaar said the methodology is being worked out.

"We will be looking at a variety of characteristics, but it is possible that an individual's cumulative forecast history will be the best predictor of his future performance," he said.

To hedge its bets, ARA is also pursuing two approaches. One is to track and weight forecasters, and the other is to use sophisticated mathematical techniques to improve aggregation of judgments instead of the usual method of simply averaging them.

Is it really feasible to identify personalities prone to making the best predictions? The answer is a definite maybe, said Dylan Evans, a lecturer in behavioral science at Ireland's University College, Cork.

"I'm not sure it's feasible to weight analysts in an aggregate forecast, but I'm not sure it's unfeasible either. It's an open question, and one which warrants further investigation of the kind that the IARPA project seems to envisage," he said.

This approach might be most fruitful for selecting the best candidates to become intelligence analysts, Evans said.

Evans, an expert on risk intelligence, said researchers have been able to identify some personality traits that correlate with erroneous forecasts, such as narcissism, extroversion and Machiavellianism. Traits associated with better forecasting have been more elusive. Intelligence and education have not been linked to more accurate predictions.

Projects like ACE suggest that the human dimension in forecasting is again taking center stage. The human role has been neglected for decades in places like Wall Street, which relied on elaborate computer models.

"An overreliance on computer models can drown out serious thinking about the big questions, such as why the financial system nearly collapsed in 2007-08, and how a repeat can be avoided," Evans said.

Researchers such as Warnaar and Evans frequently cite the work of Philip Tetlock, a psychology professor at the University of Pennsylvania, who studied the political and economic predictions of experts and found they weren't any more accurate than random chance. Tetlock is one of the leaders of the University of California, Berkeley/University of Pennsylvania ACE team, which has established the Good Judgment Project. Like the ARA effort, the Good Judgment Project is inviting members of the public to make predictions about world events. One difference is that the ARA forecasters will be anonymous to the researchers, but the Good Judgment Project will know who its forecasters are and will require that they have at least a bachelor's degree. Good Judgment forecasters will also be paid an honorarium of $150 per year.

Tetlock, author of "Expert Political Judgment: How Good Is It? How Can We Know?" used the typology of the fox and the hedgehog. The fox is the diverse thinker who knows a little about a lot of different subjects and has a modest assessment of his own capabilities, while the hedgehog is the avowed expert on a single, large subject. Tetlock found that foxes make more accurate forecasts than hedgehogs.

Does Tetlock believe it's possible to determine whether Analyst X's forecast deserves more weight than analyst Y's? The answer is yes, he said by email. "But replication and robustness tests are essential, which is another reason for the [ACE] project. We also need a better understanding of underlying mechanisms via which judgment might be helped or hurt. My favorite mediator hypothesis: the capacity for constructive self-criticism, rather than just being cautious and clinging to the midpoints of a probability scale."

Warnaar said there is a bit of fox and hedgehog in all of us. "In my view, you need both hedgehogs and foxes. Hedgehogs would be good at coming up with the subquestions, whereas foxes would be best at considering all the subquestions and providing a combined opinion on the target question."

C4ISR JOURNAL

August 2011

A GANNETT COMPANY

CYBER ALERT

Washington ratchets up preparations for the new existential threat 16

INTEL FORECASTING

Researchers ask public for help 24

MARITIME GAP

U.K. weighs C-130 decision 34

OVERSEEING INTEL

Congressman talks process, priorities 40