A conversation with Philip Tetlock, March 8, 2016

Participants

- Professor Philip Tetlock – Annenberg University Professor, University of Pennsylvania; Co-Creator, Good Judgment Project
- Holden Karnofsky – Executive Director, Open Philanthropy Project
- Luke Muehlhauser – Research Analyst, Open Philanthropy Project
- Helen Toner – Research Analyst, Open Philanthropy Project

Note: These notes were compiled by the Open Philanthropy Project and give an overview of the major points made by Professor Tetlock.

Summary

The Open Philanthropy Project spoke with Professor Tetlock of the Good Judgment Project about potential ways for the Open Philanthropy Project to improve its internal forecasting abilities. Conversation topics included the use of calibration software, question clusters, and techniques for making judgments as a group.

Calibration software

The Open Philanthropy Project has considered contracting a software developer to create a probability calibration application (for example, users could answer questions while indicating their confidence level, to evaluate and improve their calibration).

Professor Tetlock thinks such software could be a valuable tool. A software provider (formerly called Inkling, since acquired by Cultivate Labs) that works with Good Judgment may be fairly close to releasing a similar app. One of Cultivate's software tools offers the user feedback on their Brier score (a measure of the accuracy of probabilistic predictions).

It might be possible to run a private prediction challenge through Cultivate. Open Philanthropy could also potentially become a co-sponsor of the existing public platform and create a section of questions tailored to its needs (e.g. the Economist has done this).

Bayesian question clusters

Because much of the Open Philanthropy Project's grantmaking targets or depends on long-term, "scenario"-like outcomes, it might be beneficial for the Open Philanthropy Project to break down predictions about its medium- to long-term goals into "Bayesian question clusters." Open Philanthropy could make specific predictions about what it would expect to observe in the relative near-term, both within grantee programs and in relevant external circumstances and developments, if these were on the longer-term trajectory that Open Philanthropy anticipates.

For example, it could be useful to come up with a set of indicators that one would expect to observe if artificial intelligence (AI) were advancing quickly enough to significantly disrupt labor markets. Such indicators might include new AI feats (such as the recent AlphaGo victory, or the existence of driverless Ubers by the end of 2017).

Training Open Philanthropy staff to be better forecasters would not necessarily have to involve practice making a lot of predictions about, e.g., political events that aren't directly related to Open Philanthropy's goals and interests. To practice making probability judgments, Open Philanthropy could find or create a substantial set of questions that will resolve within the next six months. It might be beneficial to front-load the set with some predictions that will resolve within one month to get quicker feedback and be able to make any necessary course corrections sooner.

Question generation

In its workshops with other organizations, Good Judgment has typically found that specialist insiders in a particular field have a forecasting advantage over outside generalists. Professor Tetlock believes that domain-specific knowledge may be quite beneficial for question generation in particular. The skill set of a good forecaster may not correlate strongly with the skill set needed for good question generation.

Professor Tetlock does not think that an experienced question-generator (e.g. someone who has produced questions for forecasting tournaments) would necessarily be better than an Open Philanthropy staff member trained to do this, because it would take an outsider some time to learn relevant domain-specific knowledge within Open Philanthropy's focus areas.

Non-Open Philanthropy forecaster for benchmarking

It might be useful to use an outside forecaster for benchmarking; i.e., if Open Philanthropy's internal specialists fail to make more accurate predictions within their focus areas than smart outside generalists, it would suggest that Open Philanthropy has room to improve its forecasting significantly. Hiring an outside forecaster on a contract basis for this purpose would likely not be particularly expensive.

Group techniques

Professor Tetlock thinks techniques for aggregating team judgments are especially beneficial. In many teams, each individual contributes a set of judgments that are (by default) disconnected from and not conditioned on those of other team members. In his work with organizations, Daniel Kahneman has become particularly interested in the extent to which organizations can improve performance by reducing noise in judgments through group techniques.

For example, in addition to encouraging team members to make judgments more thoughtfully in general, sharing predictions among team members can help to reduce statistical noise. The average of several judgments is less noisy than individual judgments, with a net effect of improving accuracy. This is a relatively straightforward process, but in practice is not often done.

Delphi method

When producing predictions as a team, a potential method for avoiding complications due to, e.g., status, groupthink, reluctance to express unpopular views, etc., is to have each team member anonymously submit his or her probability judgment and explanation for it. These are then shared and discussed, and the process is iterated. Typically, judgments converge on further iterations. This process can produce about a 10% increase in predictive accuracy.

Adversarial collaboration

For questions on which there is broad disagreement (e.g. when strong AI will be developed), Professor Tetlock recommends adversarial collaboration, in which two sides with differing predictions each propose a set of resolvable indicators that they believe they have a comparative advantage in predicting. "Victory" in this exercise consists in making more accurate predictions than the opposing side about its proposed questions.

"Wisdom of crowds"

Professor Tetlock estimates that algorithmic aggregating of the predictions of a large group (e.g. 300) of typical, good-judgment forecasters can produce results nearly as accurate as a small group (e.g. 10) of superforecasters.

Temporal scope insensitivity

It is valuable for forecasters to be trained to adjust their estimates properly depending on the timeline given in the question. The typical forecaster does not strongly distinguish between the likelihood of an event happening within, e.g., three, six, or twelve months, which Daniel Kahneman terms “temporal scope insensitivity.” Professor Tetlock's superforecasters, on the other hand, tended to be moderately-to-very sensitive to temporal scope (even before awareness of temporal scope was explicitly added to the superforecaster guidelines).

Kahneman hypothesizes that this error may stem from forecasters thinking in terms of "causal propensities" – i.e. deciding how likely a given event seems given the relevant situation's current features, and simply translating that into a probability. Predictions produced in this way fail to be properly calibrated to the time dimension.

Professor Tetlock suggests *Thinking in Time* by Richard Neustadt on the psychology of this phenomenon.

Materials from Good Judgment

Professor Tetlock can provide the training materials used by the Good Judgment Project, which are now public. A randomized controlled trial of Good Judgment training showed about 10% improvement in trainees' predictive accuracy, lasting as long as four years after the training.

Good Judgment, Inc. has worked with other organizations that have wanted to improve their forecasting ability, and may be able to provide further suggestions about how to run teams (e.g. methods of aggregating group judgments, including psychological/interactive methods as well as statistical ones), as well as some simple ways to increase reliability of judgments.

Other people to talk to

- From Good Judgment, Inc.: Terry Murray (CEO) and Andrew Chiu (Senior Vice President)
- Adam Siegel (co-founder and CEO of Cultivate Labs), about calibration software
- Keith Stanovich (University of Toronto)
- Jonathan Baron, author of *Thinking and Deciding*
- Jeffrey Freidman (Assistant Professor of Government, Dartmouth College)
- Welton Chang (consultant on the Good Judgment Project and doctoral candidate at the University of Pennsylvania)
- Jason Matheny (Director, Intelligence Advanced Research Projects Activity) about RCTs on intelligence analyst training techniques
- Stewart Brand (President, Long Now Foundation) has been a major supporter of Professor Tetlock's work, and is particularly interested in bridging the gap between quantitative and scenario-type prediction. Professor Tetlock thinks Brand might be interested in collaborating with Open Philanthropy.

All Open Philanthropy Project conversations are available at
http://www.openphilanthropy.org/research/conversations