Expert Judgment in Early Warnings of Mass Violence: Extracting Useful Signals from Noisy Indicators

Philip E. Tetlock, University of Pennsylvania

This memo was adapted from a PowerPoint presentation delivered by Professor Tetlock at the U.S. Holocaust Memorial Museum's Committee on Conscience Seminar on Early Warning, October 5, 2011.

How can we utilize expert judgment to improve forecasting on mass violence? Research on "actuarial" vs. "clinical" expert judgment suggests a division of labor. We should rely on "content" specialists for raw logical inputs in forecasting and "process" specialists for formally integrating the inputs (combinatorics) and producing probabilistic forecasts.

Existing research shows that forecasting is tougher than most experts appreciate. With scattered exceptions (e.g., meteorologists, master bridge players, race handicappers), experts are quite poorly calibrated—and political experts are among the worst calibrated of all samples. My *Expert Political Judgment* (2005) project applied standard scoring rules to 27,000 probability judgments of political-economic outcomes spanning 55 countries over 20 years. 35% of experts could not out-perform a random-guess generator and 60% could not out-perform simple extrapolation algorithms. When experts were 90% confident, the outcome happened only about 68% of the time. When they were 10% confident, the outcome occurred about 30% of the time.

Experts are especially poorly calibrated in estimating rare events and high-stakes events. Experts often conflate facts and values and exaggerate events they want others to worry about. Unfortunately, orchestrated mass murders fall in both categories.

Although most experts may not be great forecasters, they do possess a lot of knowledge that, properly elicited, can become raw content inputs into forecast-generation algorithms. The challenge is unlocking this knowledge. Unlocking it requires knowing which questions to ask, how to frame them, how to elicit answers, and how to translate them into algorithm-friendly

forms. In this way, subjective inputs plus objective combinatorics get you as close as possible to the optimal forecasting frontier.

This is exactly what we are trying to do in a new forecasting tournament sponsored by the Intelligence Advanced Research Projects Agency (IARPA). First, we are creating flexible interdisciplinary teams with "content" and "process" specialists with rapid response times. "Content" specialists provide the raw inputs for generating forecasts (e.g., reference classes for "priors" and signposts for "updating") that "process" specialists integrate into forecasts. The goal is bridging the gap between abstract process prescriptions and specific forecasting problems.

It is one thing to exhort analysts to use reference classes or signposts correctly—and quite another to pull it off. The goal is not to put experts on the spot by asking for forecasts. Rather, we must seek to elicit key content inputs that process specialists can use either to generate forecasts or to correct those already made. In the IARPA project, we are now eliciting six categories of "content inputs," each of which could be potentially useful to a mass violence early-warning project. These inputs include:

- **Reference Classes and "Priors:"** Experts may not be great forecasters, but—looking back in time—they do know how to spot relevant historical analogies and reference classes that, with technical help, can be translated into realistic "priors" or initial probabilities. ***Recommendation***: Ask experts to judge the similarity of current situations to relevant reference classes of past situations in which mass murder did or did not occur.

- **Signposts and Belief Updating:** Experts may not be great forecasters, but—looking forward in time—they know what we need to know to predict outcomes of interest. These insights about "signposts" can become inputs in Bayesian belief-updating equations. ***Recommendations***: Ask experts to spot signpost indicators in the news that would induce reasonable observers to change their minds. Ask for inputs into likelihood ratios (how likely is signpost if we are or are not on historical track toward orchestrated mass violence). Use Bayesian inference software to process multiple signposts and their interdependencies.

- **Network Credibility Weights:** Experts may not be great forecasters, but they know who to ask for better-than-average forecasts. ***Recommendation***: Ask experts to nominate sources well-positioned to judge risks of mass violence and assign network-weighted credibility estimates.

- **Meta-Forecasting:** Experts may not be great forecasters, but some may be skilled at predicting the predictions other experts make on signpost indicators—and the types of mistakes they are likely to make. ***Recommendation***: Ask experts not for predictions, but rather meta-predictions on signpost indicators, for instance, how likely does a sample of experts judge various signpost indicators of risk and who in the sample is likely to under- or over-estimate the odds? In aggregating predictions, shift to weighted averaging (more weight to those good at predicting others' predictions—and their mistakes).

- **Awareness of the Boundaries of the Predictable:** Experts may not be great forecasters, but they know how to distinguish outcomes that are more or less predictable at any given juncture. ***Recommendation***: Ask experts to rate the "predictability" of questions and apply different statistical transformations to the answers to questions of varying predictability.

- **Shortcuts in Game-Theory Forecasting:** Experts may not be great forecasters, but they have a keen sense for the relative power of key players and their motivation to prevail on issues bearing on risks of mass violence. ***Recommendation***: Ask experts about the relative power and outcome preferences of key players. Compute the probability of mass killing by calculating how strongly it is preferred by more or less influential actors in the decision making process.

Conclusions

Incremental improvements in expert forecasting of mass killing are possible, but do not expect miracles. The relatively low upper limits on our ability to reliably predict outcomes like mass killing are set by the quirky path-dependent meander of history. But the opportunity costs of failing to explore even modest improvements are potentially steep in lives and money.