Obstacles to Harnessing Analytic Innovations in Foreign Policy Analysis: A Case Study of Crowdsourcing in the U.S. Intelligence Community

Laura Resnick Samotin
Columbia University

Jeffrey A. Friedman
Dartmouth College & Institute for Advanced Study in Toulouse

Michael C. Horowitz
University of Pennsylvania

October 25, 2022

Abstract

We conducted interviews with national security professionals to examine why the U.S. Intelligence Community has not systematically incorporated prediction markets or prediction polls into its intelligence reporting. This behavior is surprising since crowdsourcing platforms often generate more accurate predictions than traditional forms of intelligence analysis. Our interviews suggest that three principal barriers to adopting these platforms involved (i) bureaucratic politics, (ii) decision-makers lacking interest in probability estimates, and (iii) lack of knowledge about these platforms' ability to generate accurate predictions. Interviewees offered many actionable suggestions for addressing these challenges in future efforts to incorporate crowdsourcing platforms or other algorithmic tools into intelligence tradecraft.

Acknowledgments

Marco Allen and Sarah Turley provided excellent research assistance. Discussions with Perry World House's Predictive Intelligence Assessment Methods (PRIAM) working group played a key role in shaping the project. We thank two anonymous reviewers for offering constructive feedback. We particularly appreciate the government officials who generously provided their time and insight to make this research possible. Funding from the French Agence Nationale de la Recherche (under the Investissement d'Avenir programme, ANR-17-EURE-0010) is gratefully acknowledged. The views expressed in this article are those of the authors and do not reflect the official policy of the U.S. Government. Author names have been ordered randomly.

Word count

9,975

The Challenge of Harnessing Analytic Innovations in Foreign Policy Analysis: A Case Study of Crowdsourcing in the U.S. Intelligence Community

Political forecasting is one of the most fundamental, and most difficult, challenges of foreign policy analysis and decision-making. When leaders make foreign policy decisions, they are predicting that the benefits of their choices will outweigh their costs.[i] Yet, international politics are so complex that it is almost impossible to forecast geopolitical risks with certainty or to fully anticipate the consequences of any foreign policy.[ii] Foreign policy analysts and decision-makers should thus have strong incentives to adopt analytic tools that improve their ability to manage this challenge.

However, the U.S. Intelligence Community (IC) has struggled to systematically incorporate new methods for crowdsourcing geopolitical forecasts into its analytic products. From 2010-2020, the IC sponsored the development of two crowdsourcing platforms: the Intelligence Community Prediction Market (ICPM) and an algorithm for analyzing prediction polls developed by the Aggregative Contingent Estimation program (ACE). Both tools generated forecasts that often proved to be more accurate than traditional forms of intelligence analysis.[iii] Yet, the Intelligence Community Prediction Market and the ACE program have had negligible impact on U.S. intelligence reporting. According to the interviews described below, output from those platforms was cited in roughly two finished pieces of intelligence analysis per year, and the ICPM is no longer in operation.

What prevented the IC from systematically incorporating crowdsourced forecasts into its intelligence reporting? In addition to addressing questions about crowdsourcing methods in their own right, this question speaks to broader debates about the challenge of harnessing analytic

innovations in intelligence and national security. As artificial intelligence (AI) and machine learning capabilities become more robust, the U.S. government will have increasing incentives to implement algorithmic tools into foreign policy analysis. States such as China and Russia have already begun to incorporate cutting-edge AI methods into their national security infrastructure.[iv] Identifying and removing barriers to harnessing these kinds of analytic innovations will thus likely play a key role in helping the United States and its allies to maintain decision advantage in a competitive geopolitical environment.[v]

We address this subject in six stages. First, we provide additional background on the U.S. Intelligence Community's crowdsourcing efforts. Next, we develop five hypotheses for why foreign policy analysts might be reluctant to incorporate algorithmic tools into their political forecasts. The first hypothesis (*perceived efficacy*) is that analysts and decision-makers may not appreciate these tools' ability to generate accurate forecasts. The second hypothesis (*ease of use*) is that crowdsourcing platforms may be too difficult to employ in real-time intelligence reporting. The third hypothesis (*nontransparency*) is that analysts and decision-makers may be reluctant to base their judgments on algorithms that they do not fully understand. The fourth hypothesis (*lack of interest*) is that analysts and policymakers may not believe that accurate probability estimates play a useful role in shaping high-stakes foreign policy decisions. The last hypothesis (*bureaucratic politics*) is that intelligence analysts lack professional incentives to incorporate crowdsourced forecasts into their reports.

The article's third section describes how we evaluated our hypotheses by conducting 25 original interviews with intelligence professionals and national security officials, all of whom had first-hand experience with, or knowledge of, the Intelligence Community Prediction Market and the ACE project. We asked those analysts to tell us how (if at all) they engaged with those platforms

when producing or consuming intelligence reports, and what obstacles (if any) they saw to those platforms' implementation. The article's fourth section shows how interviewees generally agreed that the three most important barriers to implementing crowdsourced forecasts in U.S. intelligence analysis were bureaucratic politics (cited by 60 percent of participants), lack of interest (56 percent), and perceived efficacy (48 percent). By contrast, just 8 percent of interviewees cited ease of use and lack of transparency as playing prominent roles in explaining why the IC did not systematically incorporate crowdsourced forecasts into intelligence reporting. The article's fifth section describes how our interviewees offered a wide range of additional suggestions for helping the IC to implement future crowdsourcing efforts. The article's sixth section concludes.

The U.S. Intelligence Community's (Non) Use of Prediction Markets and Prediction Polls

We view crowdsourcing as a complement – not a substitute – for traditional intelligence analysis. Thus, we do not assume that foreign policy analysts and decision-makers should rely solely or even primarily on crowdsourcing platforms when generating geopolitical forecasts. The puzzle that motivates our research is instead to ask why the IC would choose to make negligible use of those platforms in its intelligence reporting, especially given those programs' demonstrated efficacy. This section illustrates that puzzle by describing the history of the Intelligence Community Prediction Market (ICPM) and the Aggregative Contingent Estimation (ACE) program.

The Intelligence Community Prediction Market

Prediction markets allow participants to trade "options" in observed events. An example would be an option that pays $1.00 if North Korea tested a nuclear weapon by a particular date. The

market price for these options reflects the event's estimated probability of occurrence. Thus, if the market price is $0.60 for an option that pays $1 if North Korea tests a nuclear weapon by a particular date, this indicates that the market believes there is roughly a sixty percent chance that this outcome will occur.[vi]

Starting in 2010, the Director of National Intelligence's office for Analytic Integrity and Standards and the Intelligence Advanced Research Projects Activity (IARPA) launched a prediction market on the U.S. Intelligence Community's classified network.[vii] The ICPM traded in points rather than in money to avoid potential conflicts of interest, but otherwise resembled prediction markets in other fields. The platform received an estimated 190,000 forecasts from 4,300 individuals.[viii] Those forecasts were subsequently found to be more accurate than predictions made in traditional intelligence reports.[ix]

Yet, the IC was unable to find a "transition partner" who would take responsibility for integrating the ICPM's output into intelligence analysis. The National Intelligence Council initially took on this role, but then handed responsibility for program management back to IARPA. IARPA's leadership then determined that the goal of implementing a mature operational system lay outside its organizational mandate to research and develop new analytic techniques.[x] Funding for the ICPM expired in 2020. Thus, after successfully developing a platform for crowdsourcing credible geopolitical forecasts, the U.S. Intelligence Community was unable to incorporate that tool into regular intelligence reporting. One interviewee estimated that the output from the ICPM was cited in just 10-20 finished intelligence estimates over the course of its lifespan.[xi]

The Aggregative Contingent Estimation program

From 2010-2015, IARPA also funded a forecasting competition called the Aggregative Contingent Estimation (ACE) program. This program recruited several teams of researchers to design algorithms for analyzing prediction polls, in which participants assign numeric probabilities to geopolitical events. Prediction polls are easier to operate than prediction markets, because they allow participants to submit predicted probabilities directly, as opposed to influencing a crowd forecast by trading options. The winning team in this competition was called the Good Judgment Project, led by Barbara Mellers and Philip Tetlock.

The Good Judgment Project's success in developing prediction polls had three primary components. First, the Good Judgment Project trained analysts to combat cognitive biases and to use tools such as base rates to make more accurate predictions.[xii] Second, the Good Judgment Project found that grouping analysts into teams improved performance.[xiii] Third, and most importantly for our purposes, the Good Judgment Project developed an algorithm for identifying highly proficient "superforecasters" and then giving those individuals' predictions more weight when extracting wisdom from crowds.[xiv]

ACE's forecasts proved to be more accurate than those generated by the Intelligence Community Prediction Market.[xv] And, given that prediction polls are easier to use than prediction markets, there is good reason to think that the ACE's methods would be a valuable resource for the U.S. Intelligence Community. However, the IC again made negligible headway implementing this tool into intelligence reporting. Though the Good Judgment Project's training module was circulated for the purposes of training analysts, none of our interviewees was aware of any attempt to incorporate output from prediction polls into finished intelligence reports.

Obstacles to implementing algorithmic forecasts in foreign policy analysis

Why did the IC not systematically incorporate crowdsourced forecasts into intelligence reporting? To develop hypotheses for answering this question, we surveyed existing scholarship on technology adoption, risk communication, and bureaucratic politics. These literatures offer five principal reasons why the IC may struggle to adopt analytic innovations such as the Intelligence Community Prediction Market and the approaches pioneered by the ACE program.

The literature on technology adoption examines the obstacles that organizations face when it comes to incorporating information technologies into regular workflows.[xvi] The two central variables in this literature are perceived efficacy and ease of use. There is nothing particularly counterintuitive about the idea that organizations are more likely to adopt tools that seem effective and easy to use. But one key insight from the technology adoption literature is that there can be substantial differences between the ways organizations perceive a new technology's efficacy and useability, and the actual properties of that tool. For example, many people initially viewed the internet as a nerdy gimmick rather than as a foundational structure for the information revolution.[xvii] It is also possible that foreign policy analysts and decision-makers were unaware of the degree to which crowdsourcing platforms could feasibly generate accurate geopolitical forecasts.

H1 (perceived efficacy): the IC did not systematically implement crowdsourced forecasts into intelligence reporting due to a lack of awareness of those tools' ability to produce accurate predictions of geopolitical events.

H2 (ease of use): the IC did not systematically implement crowdsourced forecasts into intelligence reporting due to the perceived difficulty of operation relevant platforms.

Scholarship on risk communication is also potentially relevant for understanding why the IC did not implement crowdsourced forecasts into intelligence reporting. This literature examines

why individuals vary in their receptiveness to assessments of uncertainty, particularly to the kinds of numeric probability estimates that crowdsourcing platforms generate. This scholarship shows that many people distrust the output of analytic tools that are nontransparent, in the sense that it is difficult to follow a precise chain of reasoning. Nontransparency appears to be a major obstacle to communicating information about complex topics such as public health and climate change. The models that scientists use to analyze these phenomena are generally too complex for most citizens to understand. Even if those tools have been validated by subject matter experts, many people are naturally skeptical of analyses they cannot directly follow.[xviii]

Concerns about transparency are particularly relevant to crowdsourcing, because prediction markets and prediction polls produce probability estimates without explaining how forecasters arrived at their judgments. This can make crowdsourcing platforms seems like "black boxes" when compared to traditional intelligence reporting, where analysts are generally expected to be as thorough as possible when describing the sources, methods, and assumptions that justify analytic conclusions. It is, of course, impossible to expect that any form of intelligence analysis can ever be fully transparent: one of the primary insights from cognitive psychology is that a substantial component of analytic reasoning involves subconscious thought processes that individuals cannot control or explain.[xix] Yet, it is reasonable to expect that foreign policy analysts and decision-makers will generally view crowdsourced forecasts as being less transparent than other forms of intelligence tradecraft, and this could plausibly explain why the IC was reluctant to incorporate those methods into regular reporting.

H3 (nontransparency): the IC did not systematically implement crowdsourced forecasts into intelligence reporting because analysts and consumers did not understand the methods and reasoning that those forecasts represented.

Scholarship on risk communication also shows that many people have limited interest in understanding risks' numeric magnitudes. For instance, one prominent study showed that informing homeowners about their risk of flooding had virtually no impact on their willingness to buy flood insurance.[xx] Many medical patients also appear to be relatively unresponsive to probabilistic information about the risks and benefits of various treatment options.[xxi] These findings may partly stem from low numeracy, as some people struggle to interpret the meaning of numeric percentages.[xxii] There is also substantial evidence that decision-makers' responses to risk are largely driven by emotions and normative considerations, rather than the kinds of actuarial information that numeric probabilities provide.[xxiii] This tension is sometimes described as a "rigor-relevance tradeoff": even if geopolitical forecasts are highly accurate, the information they provide may be too narrow or too technical to inform high-stakes decision-making.[xxiv] Prediction markets and prediction polls may thus not provide the kind of information that foreign policy analysts and decision-makers find meaningful when assessing uncertainty in international politics.[xxv]

H4 (lack of interest): the IC did not systematically implement crowdsourced forecasts into intelligence reporting because analysts and decision-makers do not find numeric probabilities useful for structuring discussion of foreign policy issues.

Finally, scholarship on bureaucratic politics suggests that government officials may lack professional incentives to adopt crowdsourcing platforms.[xxvi] Crowdsourcing platforms may be seen as devaluing traditional forms of subject-matter expertise.[xxvii] Since these platforms do not inherently grant more weight to the opinions of higher-ranking analysts, they may be seen as challenging status hierarchies. Proponents of crowdsourced forecasts and other algorithmic techniques may view challenging these hierarchies as a desirable outcome, as some research suggests that subject-matter expertise is overrated as an asset for geopolitical forecasting.[xxviii] But

intelligence analysts and managers who have built careers by developing reputations for subject-matter expertise may resist having their influence attenuated by the adoption of algorithmic tools. Many intelligence analysts also reportedly worry that probability estimates could receive undue weight in performance evaluations, exposing analysts to undue criticism for making apparently inaccurate judgments.[xxix] Many government agencies may also see the implementation of novel analytic tools as lying outside their traditional organizational missions.[xxx]

H5 (bureaucratic politics): the IC did not systematically implement crowdsourced forecasts into intelligence reporting because analysts lacked professional incentives to use relevant platforms.

These hypotheses are not mutually exclusive, and they each offer substantially different implications for understanding the challenges of incorporating algorithmic tools into foreign policy analysis. For example, solving problems related to perceived efficacy or nontransparency would primarily revolve around educating analysts and decision-makers about how forecasting tools function. Solving problems related to ease of use is essentially a technical matter of designing platforms that are easier to use. Solving problems related to bureaucratic incentives and lack of interest could require catalyzing broad organizational and cultural changes. Sorting through these hypotheses' relative plausibility can thus help scholars and practitioners to prioritize efforts for developing and implementing future analytic innovations. The remainder of this study draws on original interviews to provide a clearer understanding of the obstacles to systematically implementing crowdsourced forecasts into intelligence reporting, and to generate practical recommendations for tackling those challenges.

Research design

In order to understand why the U.S. Intelligence Community did not make greater use of crowdsourced forecasts, we conducted 25 structured interviews with national security professionals who had experience producing or consuming intelligence analysis.[xxxi] The median interviewee had spent 12 years working in government (average 14, standard deviation 9). Interviewees had diverse professional experience, holding positions at the State Department, the Defense Department, Congress, and the National Security Council, along with multiple military services and intelligence agencies. Interviewees' seniority within these positions ranged widely: their roles included regional analyst, military planner, Assistant Secretary, Ambassador, and National Security Advisor.

We asked each interviewee a structured list of questions. The key question for our purposes was "What do you think is the best way to make sure new initiatives [for crowdsourced forecasting] get past the R&D stage and become more permanent programs?" We categorized interviewees' responses to this question with respect to our hypotheses.[xxxii] We adopted a permissive coding rule in which responses were considered as reflecting any hypothesis which any of this study's coauthors believed was related to the interviewee's opinion. We coded the average interviewee's response as relating to 1.8 hypotheses (standard deviation 0.5).

We contacted individuals based on prior associations with geopolitical forecasting efforts, such as the ICPM and the Good Judgment Project, or connections made during the course of academic research. Almost all individuals we contacted agreed to an interview, which mitigates concerns regarding response bias.[xxxiii] Some interviewees suggested colleagues who might be interested in an interview. We contacted those individuals in a manner consistent with "snowball" interview sampling techniques.[xxxiv] Interviewees were ensured anonymity in order to maximize candor and to reduce incentives for interviewees to provide answers consistent with their organizations'

interests.[xxxv] We treat these interviews as a basis for describing degrees of agreement and disagreement among national security professionals who had first-hand knowledge about the use (and non-use) of crowdsourced forecasts in the U.S. Intelligence Community.

This empirical approach has three principal drawbacks. First, it is possible – even likely – that our interviewees' opinions are not representative of intelligence professionals as a whole. For example, the fact that our interviewees all possessed direct knowledge of crowdsourcing platforms may suggest these individuals would be more supportive of crowdsourcing than the typical intelligence professional. Yet, it is not clear how that would bias interviewees' judgments of why the IC did not make greater use of crowdsourced forecasts. The IC could potentially investigate this subject by sponsoring internal, classified research that would reflect a more representative sample of intelligence practitioners.

Another potential concern with our methodology concerns the way that we use interviews to test the relative plausibility of five predetermined hypotheses, as opposed to employing a less-structured, "bottom-up" thematic analysis of respondent opinion.[xxxvi] Our approach has the advantage of evaluating ideas that are well-grounded in theoretical scholarship from multiple disciplines. The principal downside of this approach is that it could miss relevant factors that are specific to IC crowdsourcing efforts or that our review of prior scholarship did not capture. We thus structure our empirical analysis in two parts. First, we describe how our interview responses lined up with the article's five prespecified hypotheses. Then we provide a more open-ended discussion of insights interviewees offered that our five hypotheses did not capture. The latter section raises many additional, helpful suggestions for harnessing analytic innovations in intelligence analysis, while confirming that the article's hypotheses do, in fact, capture the most

common explanations for why the IC did not integrate crowdsourced forecasts into regular reporting.

A third caveat on our methodology is that our interviews captured national security professionals' perceptions rather than precise causal inferences. Thus, while our results capture the degree of informed consensus regarding the relative importance of different barriers to implementing crowdsourced forecasts in the IC, it is always possible that this consensus is wrong or incomplete.[xxxvii] We nevertheless believe that soliciting national security professionals' perceptions are an appropriate starting point for understanding how the IC can more effectively harness analytic innovations. Choices about whether to employ analytic methodologies are ultimately made by national security professionals based on their beliefs about whether those tools are valuable. We designed our research to understand why IC personnel did not believe that crowdsourced forecasts were worth integrating into intelligence tradecraft. Changing those perceptions is a necessary first step for any future effort to incorporate crowdsourced forecasts or similar algorithmic methods into intelligence reporting.

Primary results

Table 1 shows how three categories of responses recurred most frequently across interviews. Sixty percent of interviewees cited bureaucratic politics as an obstacle to implementing crowdsourced forecasts in intelligence analysis. Fifty-six percent of interviewees cited lack of interest among decision-makers, and forty-eight percent indicated that analysts and decision-makers did not appreciate these tools' ability to generate accurate predictions of geopolitical events. By contrast, just eight percent of interviewees argued that ease of use and nontransparency played significant roles in hindering the adopting of crowdsourced forecasts by the U.S. Intelligence Community.

These findings suggest that we can divide the article's five hypotheses into two tiers of priority. The difference in response rates across these tiers is statistically significant at the $p<0.001$ level.[xxxviii]

		Hypothesis	Proportion of responses	Common themes
Higher	Priority	*Bureaucratic politics*	60%	Analysts lack professional incentives to use forecasting tools; no office is responsible for managing these tools' implementation
		Lack of interest	56%	Decisionmakers do not see numeric probability estimates as actionable; program needs a "high-level champion."
		Perceived efficacy	48%	Analysts and consumers need more information about platforms' track record.
Lower	Priority	*Ease of use*	8%	Busy analysts do not have time to experiment with new platforms.
		Lack of transparency	8%	Analysts and policymakers do not trust unfamiliar methodologies.

Table 1. *Summary of interviewee explanations for low uptake of political forecasting tools in the U.S. Intelligence Community*.

In addition to narrowing our analytic focus to a subset of concepts for understanding the limited adoption of crowdsourced forecasts in the U.S. Intelligence Community, interviewees' responses helped to clarify what those concepts meant in practice. For example, interviewees largely agreed that problems relating to lack of interest could be solved by finding a few, high-level "champions" for the use of forecasting platforms, in a manner that would not require catalyzing broad shifts in how policymakers consume information. Section 4 provides a more systematic discussion of practical suggestions interviewees offered for implementing future efforts at crowdsourced forecasting.

In the discussion that follows, interviewees' pronouns have been randomized to preserve anonymity. Their syntax has been lightly edited for clarity and to remove identifying information.

Bureaucratic politics

Sixty percent of interviewees argued that intelligence analysts lacked professional incentives to engage with crowdsourced forecasts. For example, Interviewee 24 argued that the IC's structure "siloes" analysts to focus on completing specific tasks, and thereby discourages analysts from incorporating new techniques into their work product. Interviewees 7 and 25 both explained that the IC has no agency whose institutional mission involves implementing analytic tools that apply through the Intelligence Community writ large.[xxxix] Interviewee 9 agreed that crowdsourcing platforms were unlikely to be incorporated into intelligence analysis unless that behavior was institutionalized as part of a specific office or agency's organizational mandate. "If you're in government," she argued, "you know that when everyone is responsible for something, no one is responsible." Interviewee 6 agreed that the adoption of these tools might well depend on statute

rather than persuasion. “Unless it’s in law, it’s too loose,” she explained. “The only stuff that comes out is mandated.”

Some interviewees argued that analysts may have even seen forecasting tools as being professionally harmful. Interviewee 7 characterized crowdsourcing as a “disruptive analytic method,” because it assigns equal value to the views of intelligence analysts regardless of their professional rank, and therefore challenges status hierarchies. Interviewee 13 speculated that analysts may have feared the way that quantitative forecasts could “create a system of accountability.” “I could understand where mid-level bureaucrats would be very apprehensive about this,” she said, “because it would create accounts of where they were inaccurate.” Interviewee 11 offered a related argument about how intelligence analysts often fear being second-guessed and are thus reluctant to employ new analytic techniques that could expose them to criticism. In her view, it generally takes “a real big mess or problem like a 9/11… to have people think we’ve got to do better and have other sources.” Interviewees 18 and 22 offered similar explanations of how “middle managers” in the IC were generally risk-averse when it came to analytic innovation.

Lack of interest

Fifty-six percent of interviewees explained the IC’s struggle to implement crowdsourced forecasts in a manner that reflected lack of interest, particularly on the part of foreign policy decision-makers. For example, Interviewee 11 argued that most high-stakes foreign policy choices do not depend on determining whether the chances of some outcome were twenty percent versus forty percent. In her view, there was thus limited value in using forecasting platforms to parse probability estimates. Interviewee 15 offered a similar view of how proponents of forecasting

platforms focused too much of their attention on "asking things that can be measured," rather than tackling policy-relevant questions. He thought that forecasting tools were designed to make intelligence analysis "more accurate rather than more useful."

Interviewee 16 agreed that, while proponents of forecasting techniques might benefit from sharing information about the empirical validity of their output – which relates to concerns about perceived utility – the harder and more important task was for proponents of these programs to show that their predictions were actionable for shaping policy choices. Interviewee 14 had a similar view that forecasted probabilities were too "wonky" for most consumers of intelligence. Interviewee 23 agreed that the forecasts produced by tools such as the ICPM generally seemed "interesting, but not immediately relevant." Without doing more to explain why those predictions were important for shaping choices that intelligence consumers confront daily, he argued, these forecasts would always "stay in the stack of things that a policymaker will try to get to read to stay informed but rarely do." Interviewee 25 reported that, in his experience, intelligence consumers struggled to interpret probability assessments without receiving information about the logic driving analysts' judgments, which the ICPM and the ACE program did not synthesize systematically.

In principle, lack of interest by decision-makers could constitute a large obstacle for proponents of geopolitical forecasting to overcome. Convincing a large swathe of policymakers to place a higher priority on an unfamiliar form of information would likely be a daunting task. Yet, most interviewees who cited lack of interest as a barrier to implementing crowdsourced forecasts did not express that view.

Instead, these interviewees largely agreed that the future success of forecasting platforms would depend on securing a high-level "champion" who advocated for the platforms' use. "You just need a sponsor," explained Interviewee 18. "You need a champion for the system either on the intel side

or on the consumer side – it needs to be a sufficiently high-ranking champion, like an Undersecretary of Intelligence… someone who controls money is always good." Interviewee 21 agreed that the program needed a high-level consumer of intelligence to say "I need to see the crowdsource box" to motivate analysts to provide that information. "Anything can be done in bureaucracy if you have a champion that is a senior," said Interviewee 23. "It depends on someone saying, 'this is worth doing, let's appoint someone, put this in our budget,' which would then signal that this is something of essential interest rather than a pet project…. This will come down to someone with a senior job who wants this to happen."

These views suggest that problems related to bureaucratic politics may largely reflect the lack of interest that high-ranking government officials showed in crowdsourcing platforms: if well-placed "champions" backed the adoption of crowdsourced forecasts, that could give intelligence analysts professional incentives to use those tools. For instance, Interviewee 4 claimed that placing some crowdsourced forecasts in the President's Daily Brief would likely have a major effect in encouraging the use of the tool by offices who want their work to obtain similar impact. Interviewee 25 similarly argued that proponents of crowdsourced forecasts should develop anecdotes about those tools produced accurate predictions that affected decisions made by high-level policymakers. In his view, just a few such examples would likely do more to stimulate the use of crowdsourcing platforms by intelligence analysts than any rigorous study of those tools' analytic accuracy. This perspective suggests that one of the primary reasons why analysts perceive limited professional incentive to use crowdsourced forecasts is that they are not confident that these tools actually affect foreign policy decision-making.

Perceived efficacy

Forty-eight percent of interviewees argued that geopolitical forecasting platforms would have been more widely adopted if analysts and consumers had more awareness of the tools' value. For instance, Interviewee 17 argued that "there's nothing like showing that the tool is effective to make it popular." Interviewee 24 explained that the IC may have attempted to implement crowdsourced platforms too soon, before works such as Philip Tetlock's *Superforecasting* had built awareness of those tools' ability to produce accurate predictions about world politics.

Some responses in this category focused on the importance of shifting policymaker's views about the accuracy of geopolitical forecasts. For instance, Interviewee 17 argued that many policymakers need to be convinced that the numeric probabilities that crowdsourcing platforms produce are reliable. In his experience, most intelligence consumers shared a general belief that these percentages were too "squishy" to be a basis for decision-making. Other responses focused primarily on analysts' skepticism of how well forecasting platforms predicted world events. Interviewee 10 thus indicated that analysts simply did not see "all that much added value" in consulting crowdsourced probabilities. Interviewee 20 amplified this point when explaining why he did not use forecasting platforms in his work as an analyst. "I truly did not know it was valuable at all. I saw it as people who were bored in the IC who did not have an actual job."

Ease of use & lack of transparency

Interviewees rarely attributed the slow uptake of political forecasting tools to problems relating to ease of use or lack of transparency. Only two interviewees' responses were relevant to each of these categories, respectively.

Interviewee 17 argued that forecasting tools would have been more readily adopted if they had more "user-friendly" elements: for example, by making a mobile phone application for logging

predictions and viewing output.[xl] Interviewee 20 said that he did not feel he had enough time to incorporate forecasting tools into his work as an analyst. Since this response specifically focused on time constraints, we coded it as being plausibly relevant to ease of use, because a more streamlined, intuitive platform would likely take less time to use during the workday. However, as noted earlier, the main thrust of Interviewee 20's concerns had to do with the platforms' perceived lack of value. (He was the one who said, "I truly did not know it was valuable at all.")

The response that was most relevant to concerns about transparency came from Interviewee 6. In her view, "The methodology when it's communicated vocally is really hard to trust. You can blindly trust it, but because you don't get the methodology, the tendency is to discount it." The other response we coded as relevant to lack of transparency came from Interviewee 15, who argued that it was important for forecasts to be produced by an office that was widely trusted throughout government, otherwise analysts might distrust it by virtue of where it was based within the government. Though similar concerns likely affect the value of any intelligence product, we interpreted this response as suggesting that relatively unfamiliar tools rely more heavily on the credibility of their authors.

Discussion

The low frequency of responses relating to ease of use and transparency does not mean that those issues would not shape the success of future efforts to promote crowdsourced forecasts. Had tools such as the Intelligence Community Prediction Market been more widely adopted, it is possible that our interviewees would have developed stronger views about how to make those tools easier to use or to interpret. Yet, even if that were true, it would still be consistent with finding that other problems – lack of interest, bureaucratic politics, and perceived efficacy – played more direct

roles in explaining why the IC has not systematically implemented crowdsourced forecasts into intelligence reporting.

Within this tier of "higher priority" barriers to the use of crowdsourced forecasts, the most important challenge appears to be building policymakers' interest in using the tools, particularly high-level "champions" who could advocate for crowdsourced forecasts to be placed in high-level intelligence reports. That would, in turn, give analysts professional incentives to cite those estimates in their reporting. The fact that no office or agency currently possesses a mandate to implement crowdsourcing platforms may also undermine professional incentives to work with those tools. The next section describes several additional practical insights that our interviews produced.

Suggestions for encouraging adoption of future crowdsourcing platforms

In order to solicit advice for designing future crowdsourced forecasting platforms, we posed five questions to each interviewee. Those questions involved (i) how analysts could frame geopolitical forecasts in order to maximize their credibility; (ii) whether the IC should mandate the use of crowdsourced forecasts in some situations; (iii) whether forecasting platforms would work better in classified or unclassified formats; (iv) whether allies and partners should be invited to participate in forecasting platforms; and (v) which agency would be the best institutional home for future forecasting efforts.[xli] This section describes interviewees' responses to each of those questions. It concludes by enumerating additional suggestions that interviewees offered on other subjects, including cases where interviewees identified obstacles to implementing crowdsourced forecasts that did not align with the article's five hypotheses.

Framing geopolitical forecasts

Even though interviewees did not cite nontransparency as a major obstacle to the adoption of crowdsourced forecasts in the U.S. Intelligence Community, proponents of these platforms still have an interest in making their output as credible and as digestible as possible. In order to elicit views on what that might entail, we asked each interviewee to say what kind of context or information should be provided when crowdsourced probabilities were provided in an intelligence report.

Most responses to this question involved describing methodology. For instance, Interviewee 17 said he would "like to see some explanatory box in the product that just talked about the tool and how they derived the degree of confidence that they have." Interviewee 1 said she would want to know how the information had been crowdsourced. Interviewee 4 agreed, saying she would "want to know how the platform worked" and what the methodology was. Interviewee 16 said he would want to include context on where the information used in the prediction market came from, along with assurances that the information had not been tampered with in any way. Interviewee 5 suggested including "hard data and indicators fed into the forecast," while Interviewee 18 proposed providing information about how the questions were designed to be representative of alternate futures.

Another group of interviewees suggested that crowdsourced predictions include information about forecasters themselves. Interviewee 6 said she would "want to know how many people participated and then I'd want an accurate account of who participated." Interviewee 1 and Interviewee 4 both wanted to convey who was in the "crowd" and what their subject-matter expertise entailed, especially if the forecasting platform drew participation from general public.

Interviewee 11 wanted to include track records for how well forecasts worked in previous instances, and Interviewee 20 wanted to include the evidence behind the forecasts. As noted above, Interviewee 25 thought it was crucial for crowdsourcing platforms to gather information on the reasons that analysts developed their probability assessments, rather than simply synthesizing those assessments into an aggregate prediction.

Should the IC mandate the use of geopolitical forecasts?

The previous section noted that many interviewees believed intelligence analysts had little professional incentive to work with crowdsourced forecasts (and that some analysts might even find it professionally risky to do so). The most straightforward way to incentivize analysts to use crowdsourced forecasts would be for the IC to mandate the use of such techniques when relevant data are available. This mandate would be plausibly consistent with the analytic standards that the Director of National Intelligence established with Intelligence Community Directive 203.[xlii] Those standards require the use of structured analytic techniques to reduce bias, instruct analysts to address alternative hypotheses, and offered guidelines for expressing uncertainty in clear and structured ways.[xliii] Crowdsourcing platforms help to advance each of these goals: they clearly express uncertainty in a manner structured by algorithms, and their output can serve as an independent "check" or alternative to the views developed by traditional forms of analysis.

We asked each interviewee whether analysts should be required to mention crowdsourced forecasts in their reports and to explain when they disagree with what those forecasts say. Sixty percent of interviewees supported this idea. Several interviewees highlighted that requiring the use of forecasts in intelligence reports would be especially useful in cases of disagreement, as a way of strengthening analysis and interrogating opposing views. Interviewee 13 indicated that she

favored making the use of forecasts mandatory because it would encourage highlighting disagreement in intelligence reports. Doing so would prevent analysis from getting "politicized and cherry-picked" as it passed up the chain of command. Interviewee 17 agreed, saying, "if there's a dissent, it's good to see that dissent explained out." Interviewee 4 simply argued that "I think the only way to get buy-in and investment in the tool is if they're required to use it."

Interviewee 24 disagreed with mandating the inclusion of forecasts into reports. He indicated that if there was a requirement to use forecasts, it would "incentivize checking a box" as opposed to "establishing a norm" that forecasts are useful tools for analysis. Interviewee 2 felt similarly, saying she feared that "requiring anyone to do something like this will make them not take it seriously." Interviewee 25 agreed that requiring the use of crowdsourcing platforms could stimulate organizational resistance that could undermine the tool's use in the long run.

Should future forecasting efforts include unclassified platforms?

The article's first section described how the Intelligence Community Prediction Market operated on a classified network, whereas the ACE program, administered prediction polls to participants recruited from the general public. Crowdsourcing geopolitical forecasts over unclassified systems facilitates recruiting large numbers of participants and makes it easier to distribute analytic output throughout government. However, it is also plausibly advantageous to limit the use of crowdsourcing platforms to IC personnel who have relevant professional experience. Hosting systems on classified platforms also reduces threats of manipulation by hostile actors. We therefore asked interviewees to say whether they thought future forecasting efforts should include classified systems, unclassified systems, or both. Eighteen interviewees offered an opinion on this question.

Three of these interviewees (17 percent) indicated that the IC should only maintain classified platforms. For instance, Interviewee 23 explained that he favored a classified platform because even though it would exclude individuals from the platform, an unclassified platform could cause intelligence leaks. Six interviewees (33 percent) advocated for both a classified and an unclassified prediction market to be run in tandem. For instance, Interviewee 14 felt that an open-source, unclassified platform would be better able to stimulate participation, which would increase the validity of the market. Nine interviewees (50 percent) favored an unclassified prediction market only. For instance, Interviewee 4 said she felt that the classified prediction market could be too myopic, and therefore it would be valuable to choose an unclassified option.

Should the United States collaborate with allies when crowdsourcing forecasts?

If the principal value of crowdsourcing lies with its ability to harness a large volume of diverse perspectives, then the IC might benefit from broadening the scope of these programs to include allies and partners. We asked interviewees to say whether they thought this would be a good idea. Interviewees unanimously supported asking allies and partners to collaborate on these platforms. "Yeah, why not?" replied Interviewee 3. "Forecasting works based on assumptions and indicators. Their assumptions and culture will necessarily be different and therefore important." Several other responses agreed that allied participation could be useful for harnessing "different perspectives and awareness" (Interviewee 23), or "helping to check biases" (Interviewee 20), or how "multinational links can uncover blind spots" (Interviewee 4).

Nine interviewees volunteered that would be particularly appropriate to share forecasting platforms among the Five Eyes group of Australia, Canada, New Zealand, the United Kingdom, and the United States.[xliv] Four suggested collaborating with NATO.[xlv] Interviewee 19 noted that

collaborating with allies would likely only work with an unclassified platform. Interviewee 15 suggested that it might be easier to coordinate international collaboration if the platform was hosted by a non-governmental organization.

Who should "own" crowdsourcing programs?

We asked each interviewee to say which part of the U.S. government would provide the best bureaucratic home for promoting the use of crowdsourced geopolitical forecasts. The most common answers to this question, with ten and eight votes, respectively, were the Office of the Director of National Intelligence (ODNI)[xlvi] and the National Intelligence Council (NIC).[xlvii] Interviewees typically justified these suggestions by arguing that the Intelligence Community is the natural home for geopolitical forecasting, and by saying that ODNI and the NIC are the offices within the U.S. Intelligence Community that are best positioned to lead interagency efforts. Given that the NIC withdrew from its previous attempt to implement the Intelligence Community Prediction Market, ODNI could be the logical choice to spearhead future initiatives, particularly if it added an office to its organizational chart whose mission focused on implementing analytic tools across the IC.[xlviii]

Additional suggestions

The structured nature of our interview protocol allowed us to systematically describe interviewees' responses to a range of specific questions that we identified in advance. When responding to these questions our interviewees also offered a wide range of additional suggestions for facilitating the adoption of geopolitical forecasts in the U.S. Intelligence Community. These suggestions help to supplement our empirical analysis by collecting ideas about the obstacles to

harnessing analytic innovation in foreign policy analysis that do not clearly align with the article's five hypotheses.

Several of these suggestions focused on enhancing user experience with the platforms. Interviewee 22 mentioned that agencies should be incentivized to compete on the prediction markets, measuring their performance against one another. He thought that framing forecasting efforts as friendly inter-agency competitions might encourage analyst participation. Interviewee 1 agreed, saying that agencies should offer "fun bonuses" to encourage involvement. Interviewee 19 was more amorphous in his suggestion, but he also recommended making the platforms "fun." Interviewee 25 noted that one reason why so many analysts participated in the ICPM is that they found it interesting and enjoyable to do so. These perspectives are all consistent with research showing how "gamification" can motivate user activity in analytic systems.[xlix]

Other interviewees offered advice for optimizing forecasting questions. Interviewee 17 suggested posing questions with a "hierarchy of specificity," such that some are very general and others are more specific. He felt that a diversity of questions would make the output more useful to more policymakers. Interviewee 1 said she thought questions should be derived from the National Security Strategy so that "analysts feel they're important." Interviewee 2 similarly suggested that the forecasting questions should be guided by National Security Council taskings. Interviewee 5 advocated relying on short-term questions so that analysts could quickly see results. Interviewee 4 said that analysts should be able to design at least some of the questions on the prediction market.

Interviewee 4 said that use of the platforms should be included in analyst training. Interviewee 15 argued that analysts should be trained in "how to disagree with the platforms so that they can push back, and how to be confident if they agree." This type of training could make analysts more

confident in how to deliver prediction market outputs in their reports, especially when they do not feel that the prediction market is accurate based on their own assessments.

Finally, Interviewee 25 suggested that it might have been easier for analysts to incorporate output from the ICPM into reports if the platform's output had been easier to cite in a manner that was consistent with requirements for documenting inferences in finished estimates.

Discussion and conclusion

This article examined why the U.S. Intelligence Community did not systematically incorporate crowdsourced forecasts into its reporting, despite evidence that those tools produced accurate predictions of geopolitical events. We drew on scholarship from the fields of technology adoption, risk communication, and bureaucratic politics to motivate five hypotheses for why the IC might struggle to harness analytic innovations. We conducted 25 interviews with national security professionals to assess which of those issues appear to be most relevant for explaining the negligible use of the prediction markets and prediction polls in finished intelligence estimates.

Our interviewees primarily attributed this behavior to three factors: (i) analysts lacking professional incentives to use crowdsourcing platforms, (ii) decision-makers lacking interest in numeric probability assessments, and (iii) a general lack of knowledge about crowdsourcing platforms' ability to generate reliably accurate predictions. Of these three issues, the most fundamental appears to be lack of policymaker interest. If high-level champions advocated for crowdsourced forecasts to play a role in foreign policy decision-making, that would likely provide analysts with incentives to use those tools and stimulate broader awareness of how crowdsourcing can complement traditional methods for producing estimative intelligence.

Our interviews uncovered many suggestions for improving the design and implementation of crowdsourcing platforms in the future. Generally speaking, our interviewees supported making the use of forecasting a requirement for analysts, or at least codifying the use of crowdsourcing platforms as a method for "alternative analysis" which is already mandated by Intelligence Community Directive 203. Our interviewees unanimously believed that working with allies to maximize a diversity of views within the forecasting platform could improve effectiveness and generate wider interest in the platform's outputs. Our research suggests that ODNI should have responsibility for implementing future crowdsourcing efforts, especially if it can create a new office devoted to managing analytic systems across the IC. Interviewees generally supported placing crowdsourcing efforts on unclassified networks. They also offered a range of suggestions with respect to framing forecasts, designing forecasting questions, using nonmaterial incentives to motivate participation, and making crowdsourced forecasts easier to cite in finished intelligence.

We noted that the IC could expand on our research by conducting a larger and more representative survey of intelligence analysts.[1] Future research could also gainfully examine the use of crowdsourcing platforms by other actors. For instance, studying the U.K. Intelligence Community's "Cosmic Bazaar" program could provide useful leverage for understanding how variations in program design affect implementation. Within the United States, organizations such as Good Judgment Inc. (a consulting firm that developed from the Good Judgment Project's research) and Metaculus harness crowdsourced insight for the benefit of corporate and public sector decision-making. Canvassing their experiences may help proponents of crowdsourced forecasts make their information maximally appealing to busy consumers.

Future efforts to implement crowdsourced forecasts in the IC could also draw on a broad range of research and experience that was not available at the early stages of the Intelligence Community

Prediction Market and the ACE program. For instance, scholars now know that fine-grained variations in crowdsourced probability estimates – such as the difference between 30 and 35 percent – convey reliable information.[li] Experimental evidence suggests that these details can, in fact, influence national security professionals' decisions.[lii] Philip Tetlock's book, *Superforecasting*, raised awareness of crowdsourcing methods throughout the government and the broader public.[liii] Websites such as PredictIt.org draw large-scale public participation in prediction markets. Publications such as the *Economist* now regularly cite those markets' output in political journalism. These developments have likely raised perceptions of the efficacy of crowdsourced forecasts, and they will likely make it easier to find high-level officials who could "champion" those platforms' use.

At the broadest level, our research offers conceptual and empirical foundations for understanding how national security organizations can harness analytic innovations – particularly the kinds of algorithmic techniques that will emerge from ongoing revolutions in artificial intelligence, machine learning, and other kinds of information technology. The adoption of these tools will likely face many of the same obstacles that the IC encountered when attempting to implement crowdsourcing platforms into intelligence reporting. Identifying and overcoming these obstacles will likely play a crucial role in helping the United States and its allies to maintain their edge in interstate competition.

Bibliography

Adams, D. A., R. R. Nelson, and P. A. Todd. “Perceived Usefulness, Ease of Use, and Usage of Information Technology.” *MIS Quarterly* 16, no. 2 (1992): 227-247.

Allen, J., A. A. Arechar, G. Pennycook, and D. Rand. “Scaling Up Fact-Checking Using the Wisdom of Crowds.” *Science Advances* 7, no. 36 (2021): eabf4393.

Arrow, K. J., R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviana, T. C. Schelling, R. J. Schiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, and E. Zitzewitz. “The Promise of Prediction Markets.” *Science* 320, no. 5878 (2008): 877-878.

Barnes, A. “Making Intelligence Analysis More Intelligent: Using Numeric Probabilities.” *Intelligence and National Security* 31, no. 1 (2016): 327-44.

Berry, J. M. “Validity and Reliability Issues in Elite Interviewing.” *PS: Political Science and Politics* 35, no. 4 (2002): 679-682.

Betts, R. K. “Is Strategy an Illusion?” *International Security* 25, no. 2 (2000): 5-50.

Braun, V. and V. Clarke. “Using Thematic Analysis in Psychology.” *Qualitative Research in Psychology* 3, no. 2 (2006): 77-101.

Byrd, N. “Bounded Reflectivism and Epistemic Identity.” *Metaphilosophy* 53, no. 1 (2022): 53-69.

Camerer, C. and H. Kunreuther. “Decision Processes for Low Probability Events.” *Journal of Political Analysis and Management* 8, no. 4 (1989): 565-592.

Chang, W., E. Chen, B. A. Mellers, and P. E. Tetlock. “Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments.” *Judgment and Decision Making* 11, no. 4 (2016): 509-526.

Damasio, A. *Descartes’ Error*. New York, NY: Random House, 2006.

Davis, F. D., R. P. Bagozzi, and P. R. Warshaw. “User Acceptance of Computer Technology.” *Management Science* 35, no. 8 (1989): 982-1003.

Fingar, T. *Reducing Uncertainty: Intelligence Analysis and National Security*. Stanford, CA: Stanford Security Studies, 2011.

Friedman, J. A., J. D. Baker, B. A. Mellers, P. E. Tetlock, and R. Zeckhauser. "The Value of Precision in Probability Assessment." *International Studies Quarterly* 62, no. 4 (2017): 410-422.

Friedman, J. A., J. S. Lerner, and R. Zeckhauser. "Behavioral Consequences of Probabilistic Precision." *International Organization* 71, no. 4 (2017): 803-826.

Goldstein, K. "Getting in the Door: Sampling and Completing Elite Interviews." *PS: Political Science and Politics* 35, no. 4 (2002): 669-672.

Greenstein, C. and L. Mosley. "When Talk Isn't Cheap: Opportunities and Challenges in Interview Research." In *SAGE Handbook of Research Methods in Political Science*, edited by L. Carini and R. Franzese. London: SAGE Publishing, 2020. 1167-1189.

Halperin, M. H. and P. A. Clapp. *Bureaucratic Politics and Foreign Policy*, 2nd edition. Washington, DC: Brookings University Press, 2006.

Heuer, R. *Psychology of Intelligence Analysis*. Washington, DC: Center for the Study of Intelligence.

Holden, R. J. and B. Karsh. "The Technology Acceptance Model," *Journal of Biomedical Informatics* 43, no. 1 (2010): 159-172.

Horowitz, M. C., J. Ciocca, L. Kahn, and C. Ruhl. *Keeping Score: A New Approach to Geopolitical Forecasting*. Philadelphia, PA.: Perry World House, 2021.

Horowitz, M. C., L. Kahn, and L. R. Samotin. "A Force for the Future: A High-Reward, Low-Risk Approach to AI Military Innovation." *Foreign Affairs* 101, no. 3 (2022): 157-164.

Horowitz, M. C., B. M. Stewart, D. Tingley, M. Bishop, L. R. Samotin, M. Roberts, W. Chang, B. A. Mellers, and P. E. Tetlock. "What Makes Foreign Policy Teams Tick." *Journal of Politics* 81, no. 4 (2019): 1388-1404.

Huber, O., R. Wider, and O. W. Huber. "Active Information Search and Complete Information Presentation in Naturalistic Risky Decision Tasks." *Acta Psychologica* 95, no. 1 (1997): 15-29.

Jensen, B. M., C. Whyte, and S. Cuomo. "Algorithms at War: The Promise, Peril, and Limits of Artificial Intelligence." *International Studies Review* 22, no. 3 (2020): 526-550.

Jervis, R. *System Effects: Complexity in Political and Social Life*. Princeton, NJ: Princeton University Press, 1997.

Kent, S. 'Words of Estimative Probability.' *Studies in Intelligence* 8, no. 4 (1964): 49-65.

Kunreuther, H., N. Novemsky and D. Kahneman. 'Making Low Probabilities Useful.' *Journal of Risk and Uncertainty* 23, no. 2 (2001): 103-120.

Lanir, Z. and D. Kahneman. "An Experiment in Decision Analysis in Israel in 1975." *Studies in Intelligence* 50, no. 4 (2006): 11-19.

Leech, B. L, F. R. Baumgartner, J. M. Berry, M. Hojnacki, and D. C. Kimball. 2013. "Lessons from the 'Lobbying and Policy Change' Project." In *Interview Research in Political Science*, edited by L. Mosley. Ithaca, N.Y.: Cornell University Press, 2013. 209-224.

Magat, W., K. Viscusi, and J. Huber. "Risk-Dollar Tradeoffs, Risk Perceptions, and Consumer Behavior." In *Learning About Risk,* edited by K. Viscusi and W. Magat. Cambridge, MA.: Harvard University Press, 1989. 83-97.

Marchio, J. "The Intelligence Community's Struggle to Express Analytic Uncertainty in the 1970s." *Studies in Intelligence* 58, no. 4 (2014): 31-42.

Marrin, S. "Evaluating the Quality of Intelligence Analysis: By What (Mis)Measure?" *Intelligence and National Security* 27, no. 6 (2012): 896-912.

Matthews, P. "Why Are People Skeptical about Climate Change? Some Insights from Blog Comments." *Environmental Communication* 9, no. 2 (2015): 153–68.

McHenry, J. "Three IARPA Forecasting Efforts." Presentation at the Federal Foresight Community of Interest's 18th annual meeting, (2018). At https://www.ffcoi.org/wp-content/uploads/2019/03/Three-IARPA-Forecasting-Efforts-ICPM-HFC-and-the-Geopolitical-Forescasting-Challenge_Jan-2018.pdf.

Peters, E., J. Hibbard, P. Slovic, and N. Dieckmann. "Numeracy Skill and the Communication, Comprehension, and Use of Risk-Benefit Information." *Health Affairs* 26, no. 3 (2007): 741-748.

Satopää, V. A., J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. "Combining Multiple Probability Predictions Using a Simple Logit Model," *International Journal of Forecasting* 30, no. 2 (2014): 344-356.

Sims, J. E. “Decision Advantage and the Nature of Intelligence Analysis” in *Oxford Handbook of National Security Intelligence* edited by L. Johnson. New York, NY: Oxford University Press, 2010. 389-403.

Slovic, P., M. L. Finucane, E. Peters, D. G. MacGregor. ‘Risk as Analysis and Risk as Feelings.’ *Risk Analysis* 24, no. 2 (2004): 311-322.

Snyder, Jack. ‘Richness, Rigor, and Relevance in the Study of Soviet Foreign Policy.’ *International Security* 9, no. 3 (1984-1985): 89-108.

Stastny, B. J. and P.E. Lehner. “Comparative Evaluation of the Forecast Accuracy of Analysis Reports and a Prediction Market.” *Judgment and Decision Making* 13, no. 2 (2018): 202-211.

Stoll, C. ‘Why the Web Won’t Be Nirvana.’ *Newsweek*, February 26, 1995.

Tetlock, P. E. *Expert Political Judgment*. Princeton, NJ: Princeton University Press, 2005.

Tetlock, P. E. and D. Gardner. *Superforecasting: The Art and Science of Prediction*. New York, NY: Crown, 2015.

Werbach, K. and D. Hunter. *For the Win: How Game Thinking Can Revolutionize Your Business*. Philadelphia, PA: Wharton School Press, 2020.

Whitmarsh, L. “Skepticism and Uncertainty about Climate Change: Dimensions, Determinants, and Change Over Time.” *Global Environmental Change* 21, no. 2 (2011): 690-700.

Wolfers, J. and E. Zitzewitz. “Prediction Markets.” *Journal of Economic Perspectives* 18, no. 2 (2004): 107-126.

Zegart, A. *Spies, Lies, and Algorithms: A History and Future of American Intelligence*. Princeton, NJ: Princeton University Press, 2022.

Notes

1. [i] Betts, 'Is Strategy an Illusion?'
2. [ii] Jervis, *System Effects*.
3. [iii] Stastny and Lehner, 'Comparative Evaluation of the Forecast Accuracy of Analysis Reports and a Prediction Market'; Tetlock and Gardner, *Superforecasting*, 1-24.
4. [iv] Horowitz, Kahn, and Samotin, 'A Force for the Future.'
5. [v] Sims, 'Decision Advantage and the Nature of Intelligence Analysis'; Jensen, Whyte, and Cuomo, 'Algorithms at War'; Zegart, *Spies, Lies, and Algorithms*.
6. [vi] For general background on prediction markets as a forecasting tool, see Arrow et al., 'The Promise of Prediction Markets'; Wolfers and Zitzewitz, 'Prediction Markets.'
7. [vii] Horowitz, Ciocca, Kahn, and Ruhl, *Keeping Score*.
8. [viii] McHenry, 'Three IARPA Forecasting Efforts.'
9. [ix] Stastny and Lehner, 'Comparative Evaluation of the Forecast Accuracy of Analysis Reports and a Prediction Market.'
10. [x] Interviewee 7.
11. [xi] Interviewee 25.
12. [xii] Chang, Chen, Mellers, and Tetlock, 'Developing Expert Political Judgment: The Impact of Training and Practice on Judgmental Accuracy in Geopolitical Forecasting Tournaments.'
13. [xiii] Horowitz et al., 'What Makes Foreign Policy Teams Tick.'
14. [xiv] Satopää et al., 'Combining Multiple Probability Predictions Using a Simple Logit Model.'
15. [xv] Tetlock and Gardner, *Superforecasting*: 1-24.
16. [xvi] Davis, Bagozzi, and Warshaw, 'User Acceptance of Computer Technology'; Adams, Nelson, and Todd, 'Perceived Usefulness, Ease of Use, and Usage of Information Technology'; Holden and Karsh, 'The Technology Acceptance Model.'
17. [xvii] See, for example, Stoll, 'Why the Web Won't Be Nirvana.'
18. [xviii] Matthews, 'Why Are People Skeptical about Climate Change?'; Whitmarsh, 'Skepticism and Uncertainty about Climate Change.'
19. [xix] For instance, see Damasio, *Descartes' Error*. On the relevant of this research for intelligence studies, see Heuer, *Psychology of Intelligence Analysis*.
20. [xx] Camerer and Kunreuther, 'Decision Processes for Low Probability Events'; Magat, Viscusi, and Huber, 'Risk-Dollar Tradeoffs, Risk Perceptions, and Consumer Behavior'; Huber, Wider, and Huber, 'Active Information Search and Complete Information Presentation in Naturalistic Risky Decision Tasks.'
21. [xxi] Peters, Hibbard, Slovic, and Dieckmann, 'Numeracy Skill and the Communication, Comprehension, and Use of Risk-Benefit Information.'
22. [xxii] Kunreuther, Novemsky, and Kahneman, 'Making Low Probabilities Useful'; Peters et al., 'Numeracy Skill.'
23. [xxiii] Slovic, Finucane, Peters, and MacGregor, 'Risk as Analysis and Risk as Feelings.'
24. [xxiv] Snyder, 'Richness, Rigor, and Relevance in the Study of Soviet Foreign Policy.'
25. [xxv] See, for example, Kent, 'Words of Estimative Probability'; Lanir and Kahneman, 'An Experiment in Decision Analysis in Israel in 1975'; Marchio, 'The Intelligence Community's Struggle to Express Analytic Uncertainty in the 1970s.'
26. [xxvi] Horowitz et al., *Keeping Score*, 17-18.

27. [xxvii] Indeed, the Good Judgment Project showed that crowdsourced prediction polls were generally more accurate than the judgments of any individual analyst, and that subject matter expertise had relatively little impact on the accuracy of analysts' forecasts. See Tetlock and Gardner, *Superforecasting*, 250-270.
28. [xxviii] See, for example, Tetlock, *Expert Political Judgment*. For broader evidence that the collective wisdom of amateurs can be as reliable as expert judgment, see Allen, Pennycook, and Rand, 'Scaling Up Fact-Checking Using the Wisdom of Crowds.'
29. [xxix] Barnes, 'Making Intelligence Analysis More Intelligent'; Marrin, 'Evaluating the Quality of Intelligence Analysis.'
30. [xxx] Halperin and Clapp, *Bureaucratic Politics and Foreign Policy*, 25-61.
31. [xxxi] Interviews were conducted between June 2021 and May 2022.
32. [xxxii] Thus, we treated each interview as a unit of observation with five binary variables, reflecting whether or not the interviewee's views were consistent with each hypothesis. In some cases, interviewees volunteered information about obstacles to implementing crowdsourced forecasts in response to other questions. If those responses were relevant to a hypothesis, they were also categorized for these purposes. Each member of the research team independently coded each interview. Discrepancies were identified and debated to reach unanimous judgments in coding the binary measures.
33. [xxxiii] Goldstein, *Getting in the Door*, 670.
34. [xxxiv] Interviews were conducted in-person, on zoom, and by telephone, depending on the availability, location, and preference of the interviewees. All interviews were thirty minutes, and interviewees were not compensated for their time. All interviews were conducted under the auspices of University of Pennsylvania IRB Protocol #848953.
35. [xxxv] Berry, 'Validity and Reliability Issues in Elite Interviewing.'
36. [xxxvi] See, for example, Braun and Clarke, 'Using Thematic Analysis in Psychology.'
37. [xxxvii] Greenstein and Mosley, 'When Talk Isn't Cheap'; Leech et al. 'Lessons from the 'Lobbying and Policy Change' Project,' 214.
38. [xxxviii] The difference in response proportions among the three "top tier" responses is not statistically significant. For instance, the gap between response rates pertaining to bureaucratic politics and perceived efficacy has a *p*-value of 0.48.
39. [xxxix] As noted in Section 1, organizations such as IARPA are tasked to develop analytic tools with broad applicability, but not to manage the implementation of mature operational systems.
40. [xl] Interviewee 17 noted that this idea would only work when using unclassified platforms; the next section will return to the question of whether forecasting tools would be more valuable in classified versus unclassified settings.
41. [xli] These questions were designed to address directions for future research suggested in Horowitz et al., *Keeping Score*.
42. [xlii] Available at https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf.
43. [xliii] Fingar, *Reducing Uncertainty*.
44. [xliv] Interviewees 4, 5, 7, 13, 17, 20, 21, 22, 25.
45. [xlv] Interviewees 4, 5, 7, and 10.
46. [xlvi] Interviewees 3, 4, 10, 12, 13, 20, 21, 23, 24, 25.

47. [xlvii] Interviewees 2, 4, 5, 9, 11, 17, 18, 23. Interviewee 9 specifically recommended granting program management to the National Intelligence Officer for Warning. Interviewee 2 recommended the NIC's Strategic Futures Group.
48. [xlviii] Several other intelligence organizations, such as the Open Source Center, the National Geospatial-Intelligence Agency, and the MITRE Corporation, received one vote apiece. (Interviewees 9, 12, 18). Interviewee 24 suggested locating the program in Congress, but Interviewee 23 opposed that idea. No interviewees suggested housing the program with the Department of Defense (DoD), and three expressly said that DoD would be a poor home for the project. (Interviewees 5, 7, 10).
49. Interviewees 8, 12, 18.
50. [xlix] Werbach and Hunter, *For the Win*.
51. [l] Another approach would be for the IC to analyze which of its personnel were most likely to participate in the ICPM. This would provide a behavioral measure for understanding empirical variation in which some analysts or offices were more likely than others to support crowdsourced forecasts. However, given that very few analysts ultimately cited the ICPM in finished reporting, this method would primarily serve to explain why some analysts demonstrated interest in algorithmic forecasting, which is not the same thing as understanding the conditions under which analysts would be more or less likely to integrate that methodology into intelligence tradecraft.
52. [li] Friedman, Baker, Mellers, Tetlock, and Zeckhauser, 'The Value of Precision in Probability Assessment.'
53. [lii] Friedman, Lerner, and Zeckhauser, 'Behavioral Consequences of Probabilistic Precision.'
54. [liii] Tetlock and Gardner, *Superforecasting*.