

3-29-2023

## Hybrid forecasting of geopolitical events

Daniel M. Benjamin

*Nova Southeastern University, [dbenjam1@nova.edu](mailto:dbenjam1@nova.edu)*

Fred Morstatter

*USC Information Sciences Institute*

Ali E. Abbas

*University of Southern California*

Andres Abeliuk

*University of Chile*

Pavel Atanasov

*Pytho*

*See next page for additional authors*

Follow this and additional works at: [https://nsuworks.nova.edu/hcbe\\_facarticles](https://nsuworks.nova.edu/hcbe_facarticles)



Part of the [Business Commons](#)

---

### NSUWorks Citation

Benjamin, Daniel M.; Morstatter, Fred; Abbas, Ali E.; Abeliuk, Andres; Atanasov, Pavel; Bennett, Stephen; Beger, Andreas; Birari, Saurabh; Budescu, David V.; Catasta, Michele; Ferrara, Emilio; Haravitch, Lucas; Himmelstein, Mark; Tozammel Hossain, KSM; Huang, Yuzhong; Jin, Woojeong; Joseph, Regina; Leskovec, Jure; Matsui, Akira; Singh, Amandeep; Sasic, Rok; Steyvers, Mark; Szekely, Pedro A.; Ward, Michael D.; and Galstyan, Aram, "Hybrid forecasting of geopolitical events" (2023). *HCBE Faculty Articles*. 1167.  
[https://nsuworks.nova.edu/hcbe\\_facarticles/1167](https://nsuworks.nova.edu/hcbe_facarticles/1167)

This Article is brought to you for free and open access by the H. Wayne Huizenga College of Business and Entrepreneurship Faculty Publications at NSUWorks. It has been accepted for inclusion in HCBE Faculty Articles by an authorized administrator of NSUWorks. For more information, please contact [nsuworks@nova.edu](mailto:nsuworks@nova.edu).

---

**Author(s)**

Daniel M. Benjamin, Fred Morstatter, Ali E. Abbas, Andres Abeliuk, Pavel Atanasov, Stephen Bennett, Andreas Beger, Saurabh Birari, David V. Budescu, Michele Catasta, Emilio Ferrara, Lucas Haravitch, Mark Himmelstein, KSM Tozammel Hossain, Yuzhong Huang, Woojeong Jin, Regina Joseph, Jure Leskovec, Akira Matsui, Amandeep Singh, Rok Susic, Mark Steyvers, Pedro A. Szekely, Michael D. Ward, and Aram Galstyan



## ARTICLE

# Hybrid forecasting of geopolitical events<sup>†</sup>

Daniel M. Benjamin<sup>1,2</sup> | Fred Morstatter<sup>1</sup> | Ali E. Abbas<sup>3</sup> | Andres Abeliuk<sup>1,4,5</sup> | Pavel Atanasov<sup>6</sup> | Stephen Bennett<sup>7</sup> | Andreas Beger<sup>8</sup> | Saurabh Birari<sup>1</sup> | David V. Budescu<sup>9</sup> | Michele Catasta<sup>10</sup> | Emilio Ferrara<sup>1</sup> | Lucas Haravitch<sup>3</sup> | Mark Himmelstein<sup>9</sup> | KSM Tozammel Hossain<sup>1</sup> | Yuzhong Huang<sup>1</sup> | Woojeong Jin<sup>1</sup> | Regina Joseph<sup>11</sup> | Jure Leskovec<sup>10</sup> | Akira Matsui<sup>1</sup> | Mehrnoosh Mirtaheri<sup>1</sup> | Xiang Ren<sup>1</sup> | Gleb Satyukov<sup>1</sup> | Rajiv Sethi<sup>12</sup> | Amandeep Singh<sup>1</sup> | Rok Sosic<sup>10</sup> | Mark Steyvers<sup>7</sup> | Pedro A Szekely<sup>1</sup> | Michael D. Ward<sup>8</sup> | Aram Galstyan<sup>1</sup>

<sup>1</sup>USC Information Sciences Institute, Marina del Rey, California, USA

<sup>2</sup>Nova Southeastern University, Fort Lauderdale, Florida, USA

<sup>3</sup>University of Southern California, Los Angeles, California, USA

<sup>4</sup>University of Chile, Santiago, Chile

<sup>5</sup>National Center for Artificial Intelligence (CENIA), Macul, Chile

<sup>6</sup>Pytho, LLC, New York, New York, USA

<sup>7</sup>University of California Irvine, Irvine, California, USA

<sup>8</sup>Predictive Heuristics, Seattle, Washington, USA

<sup>9</sup>Fordham University, Bronx, New York, USA

<sup>10</sup>Stanford University, Stanford, California, USA

<sup>11</sup>Sibylink, New York, New York, USA

<sup>12</sup>Barnard College, Columbia University, New York, New York, USA

## Correspondence

Aram Galstyan.

Email: [galstyan@isi.edu](mailto:galstyan@isi.edu)

<sup>†</sup>In memoriam of Michael D. Ward. This paper would not be possible without his contributions.

## Funding information

Intelligence Advanced Research Projects Activity, Grant/Award Number: 2017-17071900005

## Abstract

Sound decision-making relies on accurate prediction for tangible outcomes ranging from military conflict to disease outbreaks. To improve crowdsourced forecasting accuracy, we developed SAGE, a *hybrid forecasting system* that combines human and machine generated forecasts. The system provides a platform where users can interact with machine models and thus anchor their judgments on an objective benchmark. The system also aggregates human and machine forecasts weighting both for propinquity and based on assessed skill while adjusting for overconfidence. We present results from the Hybrid Forecasting Competition (HFC)—larger than comparable forecasting tournaments—including 1085 users forecasting 398 real-world forecasting problems over 8 months. Our main result is that the hybrid system generated more accurate forecasts compared to a

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *AI Magazine* published by Wiley Periodicals LLC on behalf of the Association for the Advancement of Artificial Intelligence.

human-only baseline, which had no machine generated predictions. We found that skilled forecasters who had access to machine-generated forecasts outperformed those who only viewed historical data. We also demonstrated the inclusion of machine-generated forecasts in our aggregation algorithms improved performance, both in terms of accuracy and scalability. This suggests that hybrid forecasting systems, which potentially require fewer human resources, can be a viable approach for maintaining a competitive level of accuracy over a larger number of forecasting questions.

## INTRODUCTION

From military conflicts to disease outbreak to economic disruption, accurate prediction is vital for sound intelligence-based decision-making. However, the problem of making accurate predictions for geopolitical events is notoriously difficult due to too much or too little data, rare event occurrences, or large levels of uncertainty. Prediction methods range from expert and/or group judgment to individual and ensembled statistical models (Zellner et al. 2022). It is often challenging to identify a consistent, superior prediction method (Meehl 1954; Zellner et al. 2022). Stakeholders confidently misidentify the benefits of competing methods, such as trusting human clinical judgment over statistical or algorithmic judgment (Dawes, Faust, and Meehl 1989). Two common forecasting methods, crowdsourcing and machine learning, have complementary strengths and competing weaknesses. Here we present a hybrid forecasting model—a system that aims to exploit the proficiencies of each while circumventing their deficiencies.

Recent forecasting tournaments such as IARPA's Aggregative Contingent Estimation (ACE) (IARPA 2011) have led to advances in crowdsourcing methods, statistical aggregation, and ultimately improvements in accuracy (Atanasov et al. 2017; Mellers et al. 2014). Crowdsourced aggregation pools a breadth of knowledge while canceling independent errors (Budescu 2006) and is most successful when individual performance can be tracked over time. However, social influences can harm opinion pools, and individual (rewards) versus group incentives can be difficult to balance. Individuals must choose to share their private information and trust others.

Advances in statistical and machine learning methods lead to accuracy gains due to their ability to handle troves of data with heterogeneous input and identify complex relationships (Ghoddusi, Creamer, and Rafizadeh 2019). Data-driven, algorithmic forecasting can be used to predict various political outcomes, such as terrorism, conflict, insurgency, and similar (Enders and Sandler 2002; Hossain et al. 2020; Pilster and Böhmelt 2014; Schrodtt 2010; Schutte

2017; Vasanthan et al. 2013). However, machine learning requires large amounts of data to be available and accessible. If data are not in a standard format, there can be large costs to preprocessing data.

Statistical models perform well under the right circumstances (Leigh and Wolfers 2006), and human crowds succeed when deftly combined (Atanasov et al. 2017). Factors like amount, availability, and structure of data determine how these methods perform (Seifert and Hadida 2013). Machine-based forecasting methods typically perform well on problems for which there is sufficient historical data, but are ill-suited to forecast rare or idiosyncratic events for which such data may not exist, or when the underlying context has changed in ways not reflected by the historical data. Machine predictions handle data in a consistent, structured manner, and avoid computational errors, like violating probability axioms (Dellermann et al. 2019).

Human analysts, on the other hand, can often accurately forecast outcomes without exclusively depending on the availability of historical data, by leveraging their domain knowledge and prior experience. Further, human expertise and domain knowledge can be valuable as inputs into machine models. These benefits are most efficient when data are sparse and/or unstructured (Dellermann et al. 2019). However, even the best analysts may not match machine performance where solid historical data are available and can be cognitively overwhelmed when addressing a large number of problems within time constraints, thereby limiting the scalability of a forecasting system that relies solely on human judgment. Unfortunately, there are few direct comparisons between models and crowds in similar settings.

Here we describe our Synergistic Anticipation of Geopolitical Events (SAGE) system, which was developed under IARPA's hybrid forecasting competition (HFC) program (IARPA 2017). The system is designed to make verifiable probabilistic predictions of outcomes from a broad set of domains, such as politics and international relations (i.e., the quantity of battle deaths or piracy in a region or attributable to a specified actor), health and disease (i.e., flu or dengue fever case counts), economics

and finance (i.e., exchange rates or oil prices), and science and nature (i.e., the number of earthquakes or cybersecurity breaches) (see Section 4.1 for an overview of the types of questions). A human–computer system can achieve “hybrid intelligence” when applied in a setting with a high degree of digitization and human expertise (Rafner et al. 2021). SAGE is a hybrid forecasting platform that allows human forecasters to combine model-based forecasts with their own judgment. The SAGE system provides forecasters automated statistical predictions and freedom to choose if and how much weight to assign to model predictions when submitting their personal forecasts since formal models can increase the skill of human judges (Rafner et al. 2021).

Our system is designed to test the conceptual hypothesis that machine model forecasts embedded in a crowdsourced forecasting platform can improve the accuracy and efficiency of established crowdsourced forecasting methods. We embed machine models in a system designed to balance (a) the diversity required to achieve the “wisdom of the crowd” by not restricting users’ responses with (b) anchoring forecasters to an impartial benchmark to minimize noise and outliers. This paper tests *how* machine models lead to improvements. We experimentally test various informational conditions to determine which type of information—historical data, model output, or interactivity with the models—leads to optimal accuracy and user engagement (see Section 3 for details). We also test if machine models improve system efficiency. We hypothesize that machine models can help increase the number of questions SAGE can answer without decreasing accuracy. We test methods of allocating a fixed number of human forecasters to questions where they are most needed.

In what follows, we test if our hybrid system can improve accuracy, engagement, or scalability compared to established crowdsourcing methods. First, we discuss relevant literature related to hybrid intelligence and scalability. Then, we describe the main components of SAGE followed by a description of HFC guidelines. Finally, we present our experimental results from a 8-month long Randomized Controlled Trial (RCT) conducted under the HFC program.

## RELATED WORKS

The main motivation behind developing a “hybrid” forecasting system is to harness the strengths of crowdsourced and statistical forecasts by combining them with machine learning models as input for both human forecasters and aggregation methods. This type of *hybrid intelligence* occurs when human and machine components each contribute to a solution that outperforms and/or is more

efficient than either source on its own (Dellermann et al. 2019; Kamar 2016). Machine models, which excel at identifying patterns from data and leveraging them for making predictions, can help human judges overcome certain errors and inconsistencies. Human experts, which do not require structured input data, are capable of ad hoc feature selection, often quicker than variables can be formalized when data sources are yet unavailable.

While there is a growing field discussing the current state of hybrid intelligence, there is limited work exploring how such systems work and in what settings they excel. To date, most work explicitly discussing hybrid intelligence is theoretical (e.g., Dellermann, Lipusch, and Ebel 2017; Rafner et al. 2021, 2022). Developing an efficient and effective hybrid system to solve complex, dynamic tasks requires a carefully designed and tested machine component, a skilled human component, and principled, dynamic methods for combining them. In the current study, we address the challenges of balancing effectiveness with flexibility. Artificial intelligence exceeds when tasks are well-defined (e.g., Dellermann et al. 2019). Machine models can underperform when tasks are loosely defined, data are sparse, or environments are complex and/or changing. A forecasting tournament provides an opportunity to collect data in a structured, yet chaotic, environment. On one hand, the general question and response format is consistent and practiced users provide consistent response data. On the other hand, it is a difficult setting to generalize because new question types, sources, and datasets could be introduced after system development.

One key limitation of the previous work on crowdsourced forecasting and hybrid intelligence is that use cases are limited and often applied to a business environment (e.g., Dellermann et al. 2019; Dellermann, Lipusch, and Ebel 2017; Rafner et al. 2022). The current study is designed to provide data-driven support for the effectiveness of a hybrid system in the geopolitical forecasting domain. While there are established methods showing how crowdsourced forecasting succeeds, there are not established methods for a hybrid forecasting system (Mellers et al. 2014). Previous crowdsourced methods rely on adjustments, such as statistical recalibration, to adjust for measurable biases in human forecasters like overconfidence (Atanasov et al. 2017). It remains an open question whether the same or new cognitive biases emerge when human users interact with machine model output. Research suggests that presenting time-series data as a forecasting aid improves individual forecasts by reducing random error (De Baets and Harvey 2018). When there are detectable trends in a time series, forecasts made while viewing graphical data are more accurate than from viewing tabular data (Harvey and Bolger 1996). In

this setting, a hybrid system must account for potential, yet unmeasured, biases to effectively combine machine models and crowd predictions.

The success of machine models is driven by complexity and volatility. When predicting on real data, machine learning models face a tradeoff between the complexity of the data and the number of model parameters required to predict accurately (Parmezan, Souza, and Batista 2019). Hence, tuning and returning become cumbersome when properties of the event or dataset change. The problem becomes more challenging when predicting several periods into the future and requires methods that produce multiple outcomes (Souhaib et al. 2012). Known (simple) statistical methods often outperform more sophisticated methods (e.g., based on deep neural networks) because real-world time-series are often nonstationary. Changes from training to testing often impede how well sophisticated models generalize (Makridakis, Spiliotis, and Assimakopoulos 2018). It is yet not known whether human judgment can help identify the shifts in time-series over time that make statistical and machine model predictions miss the mark. Further, while the main focus of the machine models considered here is on quantitative, time-series data, there is also an emerging line of work, which intends to use unstructured textual data for making predictions. For example, Hossain et al. (2022) aim to extract possible precursors of certain events from documents, while Jin et al. (2021) formulate forecasting as a Question Answering (QA) problem on an appropriately selected textual dataset.

A key aspect in achieving an efficient and effective hybrid system is how to allocate both the human and machine resources. Intelligent task allocation can bring out the best in both sources (e.g., in classification (Beck et al. 2018); in consensus (Kamar 2016). In a review of 208 articles over 50 years, task allocation is identified as one of the key issues to making hybrid system work (Janssen et al. 2019). It is challenging to allocate tasks when it is not knowable in advance at which tasks machines and humans will outperform each other. As the task becomes more difficult and the system becomes more complex, task allocation becomes more difficult (Trouille, Lintott, and Fortson 2019). Introducing machine elements into crowd systems comes with trade-offs. Misallocation can diminish engagement and shift attention away from desired tasks. The “wisdom of the crowd” effect relies on sufficient expertise and diversity of knowledge. In this setting, the introduction of statistical models could diminish diversity if human participants are too trusting in the models and do not feel empowered or motivated to add their private information into the system. Intelligent task allocation must balance finding the best individual sources for a given task with maintaining a diverse pool of knowledge.

## METHODS

### SAGE system

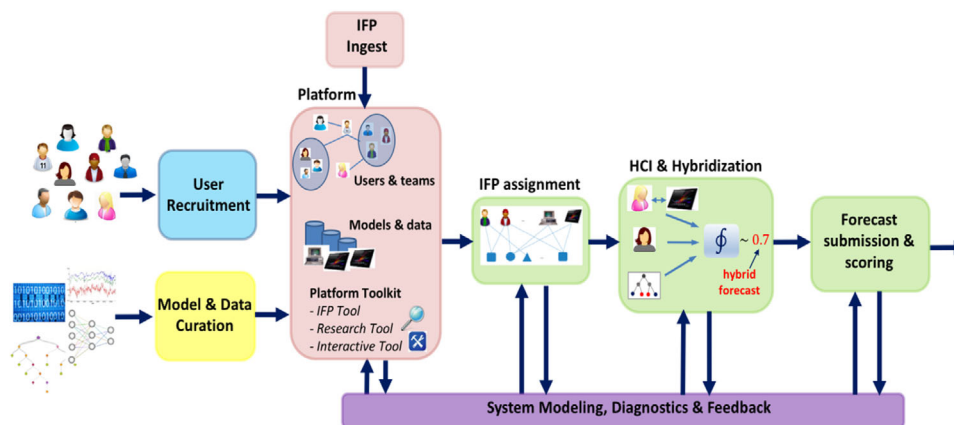
The SAGE system was developed to combine automated statistical forecasts with a pool of human knowledge by allowing users access to machine model output and by algorithmically combining human and machine forecasts (Morstatter et al. 2019). The SAGE platform allowed users to interact with machine models to anchor their judgments on an objective benchmark. Simultaneously, users had the freedom to choose if and how they combined model forecasts with their own judgment striving for the diversity of knowledge needed for the “wisdom of the crowd” effect (Abeliuk et al. 2020). To proactively mitigate skepticism with and over-reliance on the models, we trained users in how to evaluate and consolidate information from multiple sources.

The SAGE system was developed by integrating five areas of engineering and design (see Figure 1). After logging both machine and human forecasts, our aggregation algorithms computed aggregate forecasts in real time by dynamically combining human and machine forecasts. Over time, our system determined optimal weights for each source based on assessed skill, adjusting for over-confidence, and for propinquity to question resolution. We also developed a number of machine models allowing the system to choose reliable models for various question types. Finally, our system adaptively filtered questions balancing users’ preferences and abilities with the systems’ needs. Our recommender system filtered questions to the top that individuals were more likely to answer and/or were unpopular, while hiding questions that were overly popular.

### Forecasting platform

The SAGE system included search and filtering functionality to help users find forecasting questions (IFPs) about which they felt knowledgeable. At any point in time, there were dozens of IFPs available. Users had to complete at least five forecasts per week; an example of an IFP is shown in Figure 2. After choosing which question to answer, an IFP page included the following from top to bottom: question text, resolution criteria (including the source used to resolve a given question, value of interest and/or criteria for an occurrence, and timing), automated information including data graph, statistical forecast, and interactive features (depending on their experimental condition), forecast sliders that forced responses to add to 100%, a textbox to justify the





**FIGURE 1** Schematic of SAGE system organized into five topic areas. Platform engineering is in pink, recruitment and retention is in blue, machine-based forecasting is in yellow, human-machine interaction is in green, and diagnostics and feedback is in purple. SAGE, Synergistic Anticipation of Geopolitical Events.



What will be the price of gasoline in Kenya in the Nairobi market for November 2019?

**Resolution criteria:** The Famine Early Warning Systems Network (FEWS NET) provides early warning and analysis on acute food insecurity around the world. The question will be resolved using data downloaded from FEWS NET at: <http://fewsn.net/content/staple-food-price-data-1995-present>. - The relevant price can be found where 'country' is the country of interest; - 'market' is the market of interest; - 'product' is the product of interest; 'period\_date' is the date of interest; - and 'value' is the price. This question will be resolved by consulting the data source once each week sometime between 10:00 AM ET Wednesday and 4:00 PM ET Thursday, starting the Wednesday after the question close date and continuing once each week until the data are successfully retrieved.

**Category:** Politics/Intl Relations

**Opened on:** 04 September 2019

**Ending on:** 11:01 29 November 2019

[Report issues with this question](#)

**FIGURE 2** Screen capture of an IFP with resolution criteria. IFP, individual forecasting problem.

forecast, and a comments thread to view and respond to fellow users' justifications. Additional features included a leaderboard, consensus charts, a research tool, a profile page including their personal accomplishments, and training and tournament information.

## Data pipeline and model development

The SAGE machine model pipeline can be broken down into two parts based on the kinds of questions that were covered. Approximately 45% of IFPs (AKA data-driven IFPs) were clearly associated with a univariate time series, like OECD interest rates for a country (Organisation for Economic Co-operation and Development 2022). These questions were covered by automated data-acquisition and univariate time-series forecasting systems. The remaining questions did not have clearly associated data. Some, about election results and country leader resignations, were covered by tailored models that could leverage more complicated, nontime-series data. Others were covered by tools that leveraged resolved answers to other, similar, previous questions, or extracted relevant information

from the ICEWS event data (Boschee et al. 2018). The nontime-series models either only covered a very small set of questions, or did not perform well in terms of accuracy, so the rest of this section will focus on the time-series forecasting system.

The time-series forecasting system consisted of a data platform that maintained a continuously updated database of relevant time-series data sets and could map them to questions as appropriate, and a forecasting platform that would then parse a question and apply a univariate time-series model to derive probabilities for the question answers. Our system was developed to automate data extraction based on reading the question text and finding the applicable data. Many data sources were known in advance and several were not.

A significant challenge was to identify the time-series models to use for generating the forecasts that would be shown to users and sent to the aggregation models. Four core models were displayed to users in the question charts: the auto ARIMA model, a similar automated exponential smoothing model (ETS) (Hyndman and Athanasopoulos 2018; Hyndman and Khandakar 2008), a simple random walk model, and the M4-Metalearning model

(Montero-Manso et al. 2020). The DCT Ensemble model, drawing on forecasts from auto ARIMA, M4-Metalearning, or a AR(1) neural net model based on an analysis of the input series discrete cosine transformation was used to provision forecasts for aggregation. Model performance suffered from a “cold-start problem” as the number of IFP resolutions were limited over the first several months of the competition. Further, HFC guidelines required our system to make predictions for new datasets and sources on the fly, often with only a couple hours notice before users could access the IFPs. Therefore, simple time-series models tended to outperform more complex, topic-specific models—a result supported by the general success of conservative forecasting approaches (Armstrong, Green, and Graefe 2015; Makridakis, Spiliotis, and Assimakopoulos 2018). Initially, all forecasts were based on the Auto ARIMA model (Hyndman and Athanasopoulos 2018), but later, this was supplanted by an ensemble (labeled “PHE2” below) of Auto ARIMA and an exponential smoothing state space model (Hyndman and Athanasopoulos 2018), which emerged from an overall pool of 28 candidate models. We do not report results from poor performing models. Choosing adequate models was hard because interquestion performance is very noisy (variable), yet only relatively small numbers of resolved questions were available for testing and several models did not have adequate information to specify them consistently. Only later did enough resolved questions accumulate for model-to-model performance to stabilize.

## Experimental conditions

We conducted a controlled experiment to better understand the benefits of exposing forecasters to different hybridization components. We randomly assigned our 547 participants to one of three experimental conditions, which we labeled B, C, D to reflect the increasing level of complexity of and interactivity with the hybridization model.

1. **Condition B:** This condition exposed users to historical data about the target item. Data included relevant news articles from the research tool, and historical figures that pertain to the question. Historical charts were available for 177 of the 398 items.
2. **Condition C:** This condition supplemented the data charts from Condition B with machine model predictions, when available. More specifically, we exposed the forecasters to predictions from the ARIMA model, which has been determined to be a good general model. ARIMA model predictions were available for 177 of the 398 items.

3. **Condition D:** This is a variation on condition C that allows the forecasters to tweak the parameters of the visualization, including the type of model and range of data used for model training. We also provided a simple method that allowed the judges to adjust the model’s forecast by selecting the mean and variance of the target value and directly translating that into a forecast<sup>1</sup>.

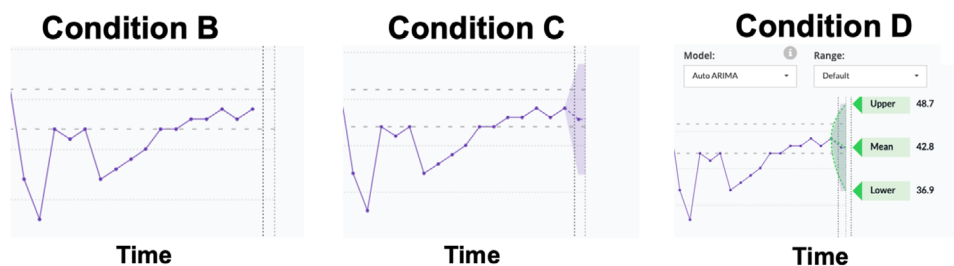
Figure 3 presents examples of screenshots from conditions B, C, and D. Control (see below) and condition B quantified the ability of human forecasters to predict the various items and provide natural baselines to compare forecasters in conditions C and D that had access to machine models. The benefit of this access is measured by the improvement in accuracy, compared to the controls. In addition to the three *treatment* conditions above, there was a control condition that was run separately by the HFC test and evaluation team. This control condition used a different platform with a different sample of 538 respondents selected from the same pool. The control condition did not offer any historical charts nor machine predictions to the participants. The main objective of the program was shown that the *hybridized* conditions could generate more accurate aggregate forecasts than the control.

## Forecast aggregation

By combining human and machine model forecasts, we aimed to leverage the collective intelligence of human and model judgments. The advantage of human judgment was its flexibility and ability to reason with qualitative and mixed-source data. Humans can forecast when data are sparse or difficult to interpret and can seek out information that only indirectly relates to the question at hand. On the other hand, the advantages of models included expeditious forecasting, which improved scalability. Statistical models dutifully forecasted on any number of questions and their accuracy tends to improve as more data became available.

The key challenge for aggregation was that many factors related to human and model judgment were not known a priori. For each particular forecasting problem, the total number of human forecasts was not knowable in advance. On any particular day the forecasting problem was available, a handful of human forecasts might be produced but in some extreme cases, no human judgment might be available for the entire duration of the forecasting problem. In addition, at the start of the forecasting project, it was not known what the relative accuracy is of the model and human judgments for certain types of forecasting problems. Every introduction of a new type of forecasting problem injected new uncertainty about the relative capabilities of human and model forecasting





**FIGURE 3** Schematic illustration of information presented to participants in each experimental condition.

accuracy. This cold-start problem made it challenging to apply machine-learning approaches that can learn optimal combinations of human and model judgment as large quantities of human judgments were initially not available and yet accurate forecasts needed to be produced from the start of the project. Therefore, the goal for aggregation was to develop a robust framework for integrating human and model judgment with the potential to scale to large numbers of forecasting questions.

To combine the human forecasts, we employed a combination of tested methods and new strategies to maximize performance. The aggregation of human-only forecasts accounted for three factors: recency, individual skill, and miscalibration. First, our algorithms diminished each forecasts' value over time as new information accumulates. To account for this recency effect, we kept only the most recent 40% of forecasts for a question at any given time, and further applied exponential decay to down-weight older forecasts included in the aggregation. Second, we placed higher weights on forecasters with better accuracy track records, those who updated their forecasts in frequent, small increments (Atanasov et al. 2020), and those who wrote longer text rationales, with more sources and quantitative information. Finally, we recalibrated forecasts to correct for the general tendency toward overconfidence by individual forecasters, and underconfidence of aggregated crowd judgments, especially when aggregated using the mean. This was done by making forecasts by individual forecasts less extreme, but aggregate-level forecasts more extreme (closer to 0 or 100%). The overall effect was to make final aggregate estimates slightly more extreme than the equivalent estimates with no recalibration. The best-performing slot used a variant of this aggregation model which made (a) forecaster weights more unequal over time, and (b) extremization parameters larger over time, making season-end aggregated forecasts more extreme than those at season-start.

Human forecasts were then combined with machine model estimates. Each model forecast was also given weights based on the historical performance of the model that generated it. Our initial strategy for human-machine aggregation was to assess machine weights relative to those

of crowd estimates (e.g., a model estimate may be weighted 1/4 as much as the crowd). The more advanced alternative that was used in most slots in the last season (including the best-performing slot), placed weights on model forecasts equal to those of several average-skill individual human forecasters. Initial human-machine weights were set based on backcasting analyses, and were allowed to vary over time based on relative within-season performance in some slots. As a consequence, the aggregate forecast was heavily weighted toward when few human forecasters had placed estimates on the question, when the human forecasts were out of date, or when most active human forecasters on the question were considered low-skill. We also tested more sophisticated ensemble aggregation methods, which used multiple machine models as inputs as well as machine learning-based aggregation, which included additional inputs such as the statistical traits of the community forecast, linguistic features of the IFPs and forecast justifications, and so forth. In a separate paper, we present a neural machine translation aggregation method, which assigns anchor attention weights to forecast-user-datetime combinations (Huang et al. 2020). We report results from the best performing aggregation method that was run in real time during the forecast season below, which is based on the method described above.

## Training

We sought to understand whether training could improve predictive accuracy under conditions in which forecasters had to balance trust in a model with their own judgment. Our HABIT training method combined probabilistic reasoning with hybridization concepts using a character-based narrative device rendered in a cartoon format. Our training was designed to extend previous vignette-based methods, which focused on core tenets of probabilistic estimation (Mellers et al. 2014), to teach about the machine models involved in the hybridized ensembles and aggregations and how to integrate model forecasts with one's personal knowledge. We hypothesized that the cognitive burden of whether to integrate or reject machine model

data could be mitigated by briefly explaining how each model works and how to balance too little and too much trust in models. We tested both whether or not mandating training improved accuracy and whether the presentation format, whether animated or static, led to gains.

## Matching participants with forecasting problems

The SAGE system aimed to optimize two seemingly conflicting objectives: (1) allow users choice of questions based on their expertise and interests with (2) timely coverage of all questions with limited human forecasters. We developed an IFP recommender system, which presented a personalized ranking of the IFPs on the question page, based on the specific characteristics of a forecaster. We develop a recommender system based on the wide and deep learning model (Cheng et al. 2016). This model identifies preferences using known IFP features, and generalizes to other IFPs via IFP embeddings. As features, we took into account the performance of the given forecaster on similar past IFPs (based on an IFP semantic similarity model we developed using BERT (Devlin et al. 2019)) and the user activity on the other IFPs (based on a collaborative filtering scheme). When designing the SAGE recommender system, BERT proved to be the more accurate, most efficient model because it captured the subtleties in the differences between IFP texts. To gauge the similarity among IFP texts, we used cosine similarity, a common distance metric used in embedding spaces. We balanced individual with system performance by also capping popular IFPs where consensus was already reached, freeing up human resources to forecast on other IFPs.

## Data availability

SAGE platform data including machine model output and experimental user forecasts, activity, and scores can be found on the Harvard Dataverse (Morstatter 2021). Control user forecasts, question metadata, and resolutions can be found on a separate Harvard Dataverse page (Hilliard 2020).

## HFC BACKGROUND AND RULES

IARPA's hybrid forecasting competition (IARPA 2017) was a multiyear research program developed to test if and how machine models could improve upon previous crowd-sourced geopolitical forecasting tournaments such as ACE (IARPA 2011). As stated in the program

announcement, “the goal of HFC was to integrate the strengths of human cognitive and reasoning abilities with those of machine-driven systems to produce maximally accurate forecasts of geopolitical and economic events” (IARPA 2017). The evaluation of hybrid forecasting system was conducted via randomized controlled trials (RCT-s). There were two RCT-s during the lifetime of the HFC program. Here we focus our analysis on the second evaluation, referred to as RCT-B, which took place from April to November 2019.

Like all forecasting tournaments, competitors must abide by rules provided by the sponsor and test and evaluation teams. Some rules governed how forecasting questions were developed, to which datasets they were linked, and how and when they would be resolved and scored. Some rules governed the activities of the human users including how they were recruited and assigned to competitor teams. Other rules governed the development and upkeep of machine-model components as well as how the machine models could be combined with human forecasts and when responses must be submitted.

## Individual forecasting problems

During RCT-B, forecasts were conducted on 398 questions, broadly referred to as individual forecasting problem (IFP). The questions covered a broad set of domains, such as politics and international relation, science, health and disease, microeconomics and finance (see Hilliard for a detailed description of different IFP types). New IFPs were published on the same day each week. Each IFP was associated with  $C$  mutually exclusive and exhaustive outcome events, where  $2 \leq C \leq 5$ . Participants submitted their forecasts for a given IFP by entering a probability for each outcome, where the probabilities across all  $C$  outcomes were required to total 100%. All IFPs had a start date and an end date during which participants could make forecasts for that questions as often as they liked (see Figure 2 for a screen capture). IFPs ranged from 2 weeks, to the full 8-month season in duration. IFPs had a mean duration of 87.07 days ( $SD = 55.85$ ). Two hundred and five IFPs had only two response options, the remaining 193 had more than two possible responses. Of these 193, 154 were ordinal, in that there was a meaningful ordering to the  $C$  events, while the remaining 39 were nominal.

## Participants

Human participants were recruited via Amazon Mechanical Turk. CloudResearch filtered the participants to ensure a high level of engagement both prior to the start of the



forecasting season by only including users with longitudinal study experience, and mid-season by removing users with low quality responses by assessing the content of their justifications (Moss 2022). The sample consisted of 547 participants, 229 women (42%), with a mean age of 36.68 (SD = 10.88). A forecasting session consisted of weekly human intelligence tasks (HITs), where each forecaster was required to make at least five forecasts. If possible, three of these five were required to be updates of previous forecasts. For each completed HIT, Participants were paid \$20 per HIT. Participants were permitted to make additional forecasts beyond these five but were not paid for these additional forecasts. Participants were also eligible for accuracy awards if they participated enough. They could earn a portion of a fixed prize pool at the midway and final points. The pool was divided among three prize tiers of \$200, \$100, or \$50 for observed accuracy—as measured with mean daily Brier (MDB) scores.

## Machine models

A key component of the HFC was the requirement for systems to produce model-based forecasts. In previous comparisons of human and model forecasts, the latter were generated in a traditional fashion by analysts (e.g., Tetlock 2017). In contrast, model-based forecasts for the HFC competition had to be generated by an automated system with restrictions on manual interventions into the process. Fixing system issues, that is, bugs and similar errors, was allowed, but manual model development like deciding what data and model(s) to use for a question, tuning model parameters, and so forth was not permitted. Some data sources were introduced mid-season. Sometimes notice about new data sources came only a couple hours prior to the associated IFPs getting published for human responses. Thus, performer teams were required to quickly produce model forecasts to aid users, and there was insufficient time to build specialized models tuned to specific datasets.

## Response submissions

Each team was allotted 40 official and up to an additional 60 experimental slots for submitting forecasts on each IFP. These slots allowed teams to test multiple theoretical ideas as well as fine-tuning the application of those ideas, such as including inputs or tuning key parameters in systematic ways. A total of 100 official slots were locked prohibiting changes by performers, and 30 were unlocked allowing for changing to key model parameters. The experimental slots encouraged testing more novel, higher risk ideas. Performer teams were required

to submit one forecast per IFP per submission slot from the day an IFP was originally published to the day it resolved, either on its stated resolution date or due to an event occurrence. For each slot, performer teams had to develop aggregation algorithms that combined the various machine and human inputs and IFP metadata for a single probability forecast.

## Scoring

The accuracy of submitted forecasts were measured using Brier scores (Brier 1950), the squared distance of the forecast from the result, coded as 1 if the event/quantity was realized, and 0 otherwise. We use Brier scores to measure accuracy because they are specifically designed to assess the accuracy of probabilistic information (unlike other metrics, like F1). As a variation of squared-error, Brier scores penalize more egregious errors more severely. In addition to scoring accuracy, a Brier score is also a proper scoring rule meaning it incentivizes responding honestly. Practically, the HFC test and evaluation team chose Brier scores as their primary accuracy metric. Using the same metric allowed us to efficiently track our performance compared to control and the other HFC competitors. Briers scores can be interpreted similarly as mean squared error. A minimal baseline for accuracy is to show improvement over an uninformed judge, who assigns equal probabilities to all  $C$  bins ( $\text{prob} = 1/C$ ), earns a Brier score of  $(C-1)/C$ . Brier scores are also commonly described as improvement over a known comparator. Below we compare to control using Cohen's  $d$ , a standardized mean difference, and in some instances display the percent improvement.

We used formulations of the Brier score based on the number of response options and ordinality of the IFP (Merkle and Hartman 2018). This Brier score variant ranged from 0 (perfect accuracy) to 2 (worst possible score). The accuracy of each forecasting slot for a given IFP was characterized by the mean daily Brier score (MDB), for example, the Brier score averaged over the active days of that IFP. Usually, the SAGE system submitted daily forecasts for each open IFP. If for whatever reason a forecast was not submitted on any given day (e.g., system outage), the last submitted forecast was carried forward. If a slot did not submit any forecasts at all for a given IFP, a uniform prior was used to calculate the score.

A similar approach was used to score individual forecasters. A forecast for a given user was carried forward until that user chose to revise the forecast. If a forecaster did not place an estimate on the first day of a question, we imputed the median score across all forecasters in a condition for each day an IFP was open prior to

**TABLE 1** The Brier scores of the best performing official and experimental methods for both the SAGE and control platforms.

Condition	Official	Experimental
Best Performing control	0.3398	0.3325
Best Performing SAGE	0.3065	0.3052

the first forecast. A user's score across IFPs was the mean of MDBs or MMDB. To adjust for the difficulty of individual IFPs and aid in interpreting comparisons across conditions, we standardized Brier scores to have a mean of zero and standard deviation of 1 for each IFP-day.

## RESULTS

### Aggregate performance

First, we report our main result that compares the aggregate performance of the SAGE system with the nonhybrid control. As we mentioned above, each method was allocated 40 official and 60 experimental slots for submitting aggregated forecasts. Table 1 summarizes our results for both official and experimental slots across 398 IFPs. The best official SAGE method led to an improvement in mean accuracy of a Cohen's *d* of 0.126 over control.

SAGE's best-performing aggregation slot had the following properties. First, it used both control and SAGE human forecaster data as inputs. Second, it applied a time-varying weighted mean human aggregation algorithm, which made forecaster weights more unequal over time; aggregate forecasters were de-extremized at the start of the season and extremization parameter value increased over time, resulting in light extremization by the end of the season. Third, human and machine-model forecasts were combined using a rule that produced a weight of a machine model forecast as equivalent to eight average-skill human forecasters on time-series questions, and four average-skill human forecasters on other questions that used less sophisticated models.

SAGE outperformed the control condition both for the official and experimental slots. We ran backcasting analyses to estimate the impact that different aspects of our system, including SAGE forecasters, human aggregation, and machine forecasts, contributed to SAGE outperforming the best control method, a Brier score difference of 0.0333. Results showed that applying the SAGE human-aggregation algorithm to control human forecasts would have resulted in a Brier score advantage of 0.01, approximately 31% of the full difference. In retrospect, the distinguishing feature of the best-performing human aggregation model was that it extremized aggregate forecasts

**TABLE 2** Number of unique forecasters and generated forecasts in each experimental condition.

Condition	Users	Forecasts	Forecasts/User
B	190	25,163	132.4
C	158	20,782	131.5
D	199	27,348	137.4
Total (SAGE)	547	73,293	134.0
A (control)	538	79,611	148.0

**TABLE 3** Comparison of standardized (at IFP level) Brier scores across conditions including mean, median, and 25% percentile for each.

Condition	Mean	Std. dev.	25th percentile	Median
A (Control)	<b>−0.021</b>	<b>0.958</b>	−0.486	<b>−0.259</b>
B (Data)	0.051	0.987	−0.472	−0.156
C (Models)	0.039	1.143	<b>−0.537</b>	−0.198
D (Interactive)	0.002	0.989	−0.508	−0.196

Bolded values represent the lowest (most accurate) score within each column.

less, especially at the start of the season. Applying this human-aggregation algorithm to the combination of control and SAGE human forecasts resulted in a Brier score advantage for SAGE of 0.032, approximately 97% of the full difference. The further addition of machine models at the aggregation stage rounded up the full 100% advantage. For more details on the accuracy benefits of machine models at the user interface versus aggregation stage (see Section 5.4).

### Individual performance across conditions

We analyzed user performance in the various conditions across the 398 resolved questions. Table 2 lists the volume of users and forecasts in each condition.

Since some questions were relatively easy and highly predictable and others were more difficult, we expected them to yield (possibly, very) different Brier scores. Thus, whenever comparing, or aggregating, Brier scores across multiple items, it was important to adjust for inherent imbalance in difficulty. Our approach to this problem was to standardize the Brier scores for every question to have a mean of 0 and a SD of 1, across all the responses in all conditions, before combining them. Thus, we report results in terms of mean, median, and 25th percentile standardized Brier scores. The lower (and more negative) a score is, the more accurate it is. The results of all conditions are presented in Table 3.

Our results indicated that users in conditions C and D outperform those in condition B, but only the most skilled forecasters outperformed the control (no data) condition.



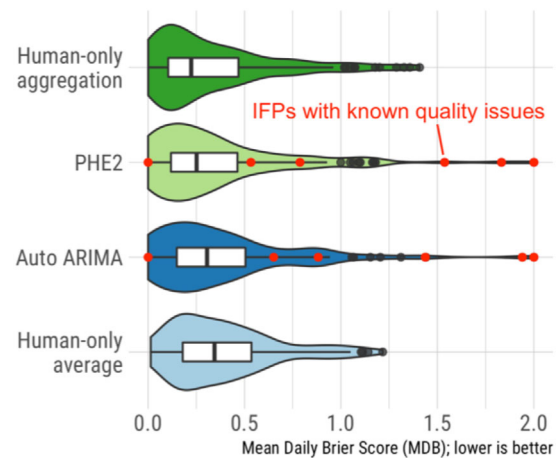
We confirmed the hypothesis that having access to model predictions indeed helps *skilled*, but not *average*, forecasters. The greatest improvement came when data charts were available, and skilled forecasters viewed model predictions,  $z\text{Brier} = -0.618$  versus  $-0.545$  for control. Note that the availability of more models and interactive features, provided in condition D, did not necessarily help with performance. Indeed, while condition D had a better mean score across all the questions, users in this condition did not perform well on questions where data charts were available. This suggests that the availability of multiple models and/or interactive features do not help the users to generate more accurate forecasts. In fact, users used the various options available to them very rarely.

Here we present overall performance of our system and experimental conditions. For an analysis of individual conditions, see Himmelstein, Atanasov, and Budescu (2021). We only find consistent differences by major IFP format, but not by traits like topic area, region, or question duration. Our model was most accurate for binary questions,  $z\text{Brier} = -0.107$  ( $SD = 1.10$ ), and least accurate for nonordinal IFPs,  $z\text{Brier} = 0.180$  ( $SD = 0.97$ ). As discussed in Section 3.5, we did include certain IFP and linguistic traits when testing more complex aggregation methods, but in most instances, they underperformed our interpretable aggregation methods except in the anchor attention model (Huang et al. 2020).

## Model-based forecasts

We also analyzed the performance of the machine models outlined in Section 3.3. Overall, simple ensemble models worked well. The best performing model (“PHE2”) was an ensemble that averaged the forecasts from Auto ARIMA and an exponential smoothing state-space model (ETS) (Hyndman and Athanasopoulos 2018). It slightly outperformed the M4-Meta model Montero-Manso et al. (2020) that ranked 2nd highest in the M4 time series forecasting competition, and clearly outperformed more complex methods like a recurrent neural network and custom-coded regularized auto-regressive model. Even the auto ARIMA model itself did reasonably well throughout. From a practical standpoint, the simpler ensembles were computationally less expensive, had fewer software dependencies, and were less likely to break.

Model forecasts relative to the human forecasts were overall near or at parity with human forecasts. Figure 4 shows the distributions of mean daily Brier scores for the auto ARIMA and PHE2 models, as well as a simple average of human forecasts and the best-performing aggregation model of human forecasters from condition



**FIGURE 4** Relative performance of two model-based forecasts compared to average human performance and the best human forecast-only aggregation model. Auto ARIMA was a mainstay model throughout; PHE2 emerged later as a top performer. This figure includes performance on 153 IFPs for which all models had forecasts. Red points mark IFPs with known quality issues that were retained for the sake of coverage. IFP, individual forecasting problem.

B, which were not exposed to machine model predictions. Both models outperformed the average human forecast but lagged slightly behind the best aggregation model of human-only forecasts. In part, this is because they had a small number of very bad forecasts. Some of these were caused by IFPs with known data quality issues, which tended to lead to extreme forecasts with either very low or very high Brier scores. Some of these—the more easily identifiable ones—are marked with the red points. Last, despite similar average performance, the model and human forecasts did well or poorly on different questions. For example, the interquestion correlation of performance for the PHE2 and aggregation benchmark models was only 0.3.

After enough IFP results were observed, we developed a meta-model for the relative performance of the time-series model and human forecasts. Overall, there were no clear bivariate or multivariate relationships between a large variety of question and data features and the relative model to human forecast performance, which for example could have led us to identify a subset of IFPs in which one consistently outperformed the other. However, towards the end of our experiment, enough performance data had accumulated so that a random forest model trying to identify forecasts that were clearly worse than a uniform forecast, or forecasts that can beat the aggregation benchmark, achieved slightly informative accuracy levels, with out-of-sample AUC-ROC values of 0.69 and 0.60, respectively. This may have been sufficient to implement a filter for likely bad forecasts, if the experiment continued.



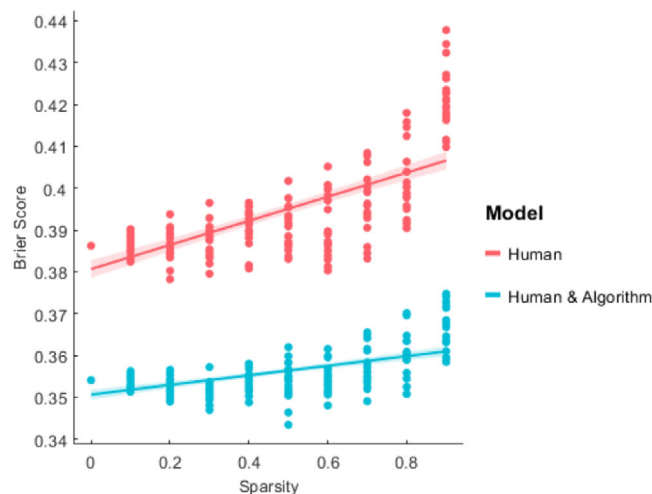
## Using machine models for scalable forecasting

We analyzed aggregate performance on the IFPs for which our strongest machine model was available, a discrete-cosine transform (DCT) ensemble. We found incorporating machine models during aggregation led to improvements at several stages, which accounted for our team's overall advantage. Although these accuracy gains were consistent throughout the competition, effects at individual stages were modest and none were statistically significant on their own. The results showed that providing forecasters access to model projections led to modest improvements in aggregate accuracy. Forecasters who could view model forecasts before making their estimates produced aggregate forecasts with 6% better Brier scores, compared to aggregations of forecasters with no access to model projections. Injecting model estimates at the aggregation stage also led to small improvements in accuracy (i.e., reductions in Brier score) of 2%–3% points.

The benefits of including model forecasts in the aggregate became more salient when considering the issue of scale. Scaling up the number of questions in a human-only pool means fewer human forecasts for each question, which is expected to degrade aggregate performance. To simulate this effect, we made human forecasts sparser by deleting a random subset of users from the aggregate. The results demonstrated that including model forecasts insulated the aggregate forecast against the negative effects of sparse human judgments that would occur when scaling to large numbers of questions (see Figure 5).

### Impact of training

Participants were randomly assigned to either a brief (about 30 min) training or a control condition in which they read popular articles about forecasting, but without tips for boosting accuracy. Training occurred once, during their second active week, and was accessible for review on the platform menu. We assessed accuracy and activity to see if trained forecasters worked harder than untrained users. We first compared forecast accuracy before and after training exposure to ensure that trained forecasters were not randomly better from the start. We found trained forecasters outperformed control forecasters post-exposure ( $d = 0.56$ ). We further assessed whether accuracy could be improved via the delivery method of the training material. Forecasters who saw animated material significantly outperformed forecasters who saw static material in average accuracy ( $t(410) = 3.55$ ,  $p < .001$ ), generated slightly more forecasts per IFP than the static group, ( $d = 0.20$ ,  $t(342) = 1.98$ ,  $p = .049$ ), and attempted approximately



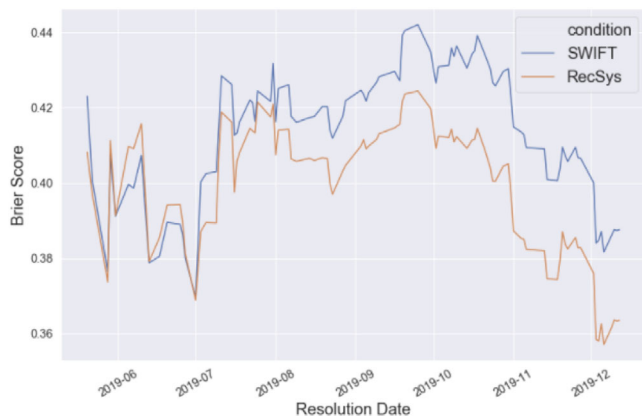
**FIGURE 5** Average aggregate performance (Brier score) as a function of the proportion of human forecasts removed from the forecasting pool (Sparsity). Higher Brier scores correspond to worse aggregate accuracy. Each point corresponds to aggregate performance for a random subset of censored forecasts. The line plots the linear regression of these points and the shaded region is the 95% confidence interval based on  $N = 20$  simulations.

5% more IFPs than static-trained counterparts, a significant difference ( $d = 0.23$ ,  $t(342) = 1.98$ ,  $p = .018$ ). More details can be found in Joseph and Atanasov (2019).

### IFP recommendation

We measured the performance benefit of users assigned to questions ordered using the IFP recommender system, described in Section 3.7, versus the global ranking based on resolution date and popularity (SWIFT ordering). As shown in Figure 6, we obtained a statistically significant improvement, up to a 7% relative decrease in Brier score by the end of the experimental phase.

We further assessed the IFP recommender system in terms of effective resource allocation by simulating different allocation strategies. First, since we could not foresee who the best performers would be before the phase ends, we implemented a greedy approach to improve our cohort of users by periodically excluding a certain percentage of worst performers every time a batch of IFP closed (i.e., got resolved) (termed **GreedyIFP**). Second, we additionally capped the number of forecasts on popular IFPs to reallocate forecasts of diminishing return after consensus was reached. We compared these methods using simple averaging of the human forecasts, omitting our aggregation algorithms and machine forecasts, to control for effects beyond resource allocation (termed **GreedyIFP++**). As shown in Table 4, both greedy strategies obtained a small improvement in the global Brier score, while reducing



**FIGURE 6** The IFP recommender system learns over time the skills and preferences of each forecaster. After 2 months into the experiment, the forecasters who received the recommendations start to consistently outperform the forecasters in the control condition (SWIFT), with a relative improvement that stabilizes around 7%. IFP, individual forecasting problem.

**TABLE 4** Brier score of different IFP recommender systems. All results are averaged over 10 runs.

Method	Brier score	Budget (% of forecasts)
All forecasts	0.397	100
Random forecasts	0.402	62
GreedyIFP	0.376	61.4
GreedyIFP++	0.375	37.8

the forecasting budget, respectively, by 38 and 62%. In contrast, we showed naively decreasing the number of forecasts (“Random forecasts”) negatively affected the global Brier Score.

## DISCUSSION

In the above, we show how a hybrid forecasting system can outperform established crowd-sourced forecasting systems. The SAGE hybrid system consistently outperforms the human-only control condition. Our hybrid system improves the accuracy in the aggregate, although improvements were modest. Beyond the proven methods for aggregating human forecasts (Atanasov et al. 2017), we find the inclusion of machine models in the user interface and the aggregation algorithms was key to improving accuracy. Individually, access to model predictions only improves the accuracy of highly skilled forecasters. While this is evidence of their value, it provides further evidence that forecasters must have enough expertise to know when and how to use this information (Abeliuk et al. 2020). Predicting the future is difficult, especially for deeply uncertain, impactful geopolitical events. Humans and machines are both limited by the irreducible uncertainty of the setting.

Combining human and machine predictions leads to gains in accuracy by helping protect against some of the most egregious errors, especially when the two sources disagree.

In addition to better forecasting accuracy, another critical advantage of SAGE over crowdsourcing-based systems is its scalability. We find evidence that the SAGE hybrid system helps answer more questions with the same number of human users without losing accuracy, although the scope of these improvements remains an open hypothesis. Adaptive question-user assignment increases the ability to scale by limiting the number of users who can access each IFP once consensus is achieved. Our recommender system succeeds based on three primary features. First, it provides a unique question ranking for each user based on the IFPs they previously chose to answer. Second, it excludes users who tend to forecast early and perform poorly. Third, it identifies an optimized number of users per IFP and capped each IFP that exceeded that maximum, thus shifting forecasts from popular to unpopular IFPs and increasing the utility of post-consensus forecasts.

Our results are subject to a number limitations, most of which are a function of participating in a forecasting tournament managed by a third-party *test and evaluation* team. Notably, our machine models needed to be robust and flexible since new datasets and question formats were regularly published unannounced. Our system was successfully able to ingest large amounts of, often unformatted, data in the window between question publication and user recruitment—often just a few hours. Thus, our results might not generalize to more stable situations with highly developed models tuned to a specific environment or dataset.

We highlight two keys lessons about the model contribution to the “hybrid” system. First, aim for depth, not breadth. Our initial strategy, in the spirit of the hybrid part of the competition, was to try to cover as many IFPs as possible. This led to several decisions to use marginal or noncanonical data. The quality issues tended to lead to extreme forecasts that sometimes were really good (achieved a low Brier score), but more often were really bad and thus reduced average quality. A better strategy would have been to focus on a smaller subset of questions where good performance can be achieved, and spend more time on quality control rather than coverage. Better average quality also simplifies downstream use of forecasts in aggregation.

Second, data is king (or more familiarly, “garbage in, garbage out”). The main cause of poor model forecasts was data quality issues. Some sources, like the OECD Organisation for Economic Co-operation and Development (2022) and OPEC Organization of the Petroleum Exporting Countries (2022), alter historic data values when updating. There were also many questions that required data transformations or had marginal data ill-suited for

time series. For example, count series based on transformed ACLED event data (Armed Conflict Location and Event Data 2022) were plagued by inaccuracies due to idiosyncratic technical issues in the data platform back-end. Unlike errors at the modeling stage, these kinds of data problems usually were hard to identify without labor-intensive manual reconstruction of a series from its source. Selecting good-enough time series models, which had been the focus of our efforts in the beginning, in the end turned to be easier than these two issues.

Similarly, the success of our simpler, interpretable aggregation methods is most likely due to model training. The greater complexity of the aggregation method, the more training data it required. Complex aggregation methods suffered from training during burn-in due to limitations in the number of IFPs that resolved earlier in the season and variability between sources in format, frequency, and availability of data. Such methods suffered more inefficiency due to retraining mid-season due to the requirement that our system be able to handle newly introduced data sources on the fly. In complex settings, simpler, traditional statistical methods often outperform novel, complex methods (Makridakis, Spiliotis, and Assimakopoulos 2018). Research on forecast combinations supports the success of simple conservative methods (Armstrong, Green, and Graefe 2015).

The recruitment of human forecasters for such long-term engagement is also innately challenging. Recruitment was managed by a third party according to HFC rules. This system prioritized retention over accuracy incentives, and HFC rules limited our ability to add performance-based incentives beyond those offered by the recruitment team. Changing the retention-accuracy incentive balance is likely to alter the quality of human performance.

In conclusion, gains from hybridizing are consistent, but modest in this setting. The SAGE system's success relies on both computer-in-the-loop hybridization including the information (historical data and model predictions) shown to users and mandated narrative graphical training, as well as human-in-the-loop hybridization including human inputs into the aggregation algorithms and strategic user-IFP assignment, to name a few. It is important to engineer such a complex system to optimize the interactions between each component, since each improves accuracy slightly. The optimal system must balance several tradeoffs, like using models that are no more complex than the tuning parameters that can be confidently estimated, and anchoring users on objective, data-driven benchmarks while eliciting the diversity required for crowd wisdom. The real advantage is not boundless improvements in accuracy. Instead it is the ability to tackle a greater burden without needing to increase human resources. When human-question balance is sparse, it is important to view users as a labor

pool and use adaptive question assignment to maximize human coverage.

## ACKNOWLEDGMENTS

The authors would like to thank Seth Goldstein, Peter Haglich, Rob Hartman, Daniel Horn, and Steven Rieber for their helpful feedback during the HFC program. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

## CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict.

## ORCID

Aram Galstyan  <https://orcid.org/0000-0003-4215-0886>

## ENDNOTE

<sup>1</sup>Behavioral decision making researchers have repeatedly documented a pattern of “Algorithm Aversion (AA for short)” (e.g., Burton, Stein, and Jensen 2020; Dietvorst, Simmons, and Massey 2015)—the tendency of humans to prefer and value advice and information from human sources over machine counterparts, even when the information provided by humans and algorithms is identical (e.g., Dietvorst, Simmons, and Massey 2015; Önköl et al. 2009). In general, judges tend to be less tolerant of errors made by algorithms, compared to humans (e.g., Dietvorst, Simmons, and Massey 2015; Pohl and Swol 2017). One way to reduce AA is to allow people to have more control over the algorithm by tweaking it, or some of its predictions (Dietvorst, Simmons, and Massey 2018). Condition D was implemented to test this expectation.

## REFERENCES

- Abeliuk, Andrés, Daniel M. Benjamin, Fred Morstatter, and Aram Galstyan. 2020. “Quantifying Machine Influence over Human Forecasters.” *Scientific Reports* 10(1): 15940.
- Armed Conflict Location and Event Data. 2022. ACLED Curated Data. June. <https://acleddata.com/curated-data-files/>
- Scott Armstrong, J., Kesten C. Green, and Andreas Graefe. 2015. “Golden Rule of Forecasting: Be Conservative.” *Journal of Business Research* 68(8): 1717–31.
- Atanasov, Pavel, Phillip Rescobar, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. 2017. “Distilling the Wisdom of Crowds: Prediction Markets vs. Prediction Polls.” *Management Science* 63(3): 691–706.
- Atanasov, Pavel, Jens Witkowski, Lyle Ungar, Barbara Mellers, and Philip Tetlock. 2020. “Small Steps to Accuracy: Incremental Belief Updaters are Better Forecasters.” *Organizational Behavior and Human Decision Processes* 160: 19–35.



- R Beck, Melanie, Claudia Scarlata, Lucy F Fortson, Chris J Lintott, B D Simmons, Melanie A Galloway, Kyle W Willett, et al. 2018. "Integrating Human and Machine Intelligence in Galaxy Morphology Classification Tasks." *Monthly Notices of the Royal Astronomical Society* 476(4): 5516–34.
- Ben Taieb, Souhaib, Gianluca Bontempi, Amir F. Atiya, and Antti Sorjamaa. 2012. "A Review and Comparison of Strategies for Multi-Step Ahead Time Series Forecasting Based on the NN5 Forecasting Competition." *Expert Systems with Applications* 39(8): 7067–83.
- Bosch, Elizabeth, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, and James Starz. 2018. ICEWS Weekly Event Data. <https://doi.org/10.7910/DVN/QI2T9A>, Harvard Dataverse, V373.
- Brier, G. W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78: 1–3.
- Budescu, David V. 2006. "Confidence in aggregation of opinions from multiple sources." In *Information Sampling and Adaptive Cognition*, edited by K Fiedler and P Juslin, 327–52. Cambridge, UK: Cambridge University Press.
- Burton, Jason W., Mari-Klara Stein, and Tina Blegind Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." *Journal of Behavioral Decision Making* 33(2): 220–39. [eprint: https://doi.org/10.1002/bdm.2155](https://doi.org/10.1002/bdm.2155)
- Cheng, Heng-Tze, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, et al. 2016. "Wide & deep learning for recommender systems." In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems (DLRS 2016)*. Association for Computing Machinery, New York, NY, USA, 710. <https://doi.org/10.1145/2988450.2988454>
- Dawes, Robyn M., David Faust, and Paul E. Meehl. 1989. "Clinical Versus Actuarial Judgment." *Science* 243(4899): 1668–74.
- De Baets, Shari, and Nigel Harvey. 2018. "Forecasting from Time Series Subject to Sporadic Perturbations: Effectiveness of Different Types of Forecasting Support." *International Journal of Forecasting* 34(2): 163–80.
- Dellermann, Dominik, Nikolaus Lipusch, and Philipp Ebel. 2017. "Developing design principles for a crowd-based business model validation system." In *Designing the Digital Transformation. Lecture Notes in Computer Science*, edited by Alexander Maedche, Jan vom Brocke, and Alan Hevner, 163–78. Springer International Publishing.
- Dellermann, Dominik, Nikolaus Lipusch, Philipp Ebel, and Jan Marco Leimeister. 2019. "Design Principles for a Hybrid Intelligence Decision Support System for Business Model Validation." *Electronic Markets* 29(3): 423–41.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1, 4171–86.
- Dietvorst, B. J., Joseph P. Simmons, and C. Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them err." *Journal of Experimental Psychology: General* 144(1): 114–26.
- Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey. 2018. "Overcoming Algorithm Aversion: People will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them." *Management Science* 64(3): 1155–70.
- Enders, Walter, and Todd Sandler. 2002. "Patterns of Transnational Terrorism, 1970–1999: Alternative Time-Series Estimates." *International Studies Quarterly* 46(2): 145–65.
- Ghoddusi, Hamed, Germán G. Creamer, and Nima Rafizadeh. 2019. "Machine Learning in Energy Economics and Finance: A Review." *Energy Economics* 81: 709–27.
- Harvey, Nigel, and Fergus Bolger. 1996. "Graphs Versus Tables: Effects of Data Presentation Format on Judgemental Forecasting." *International Journal of Forecasting* 12(1): 119–37.
- Hilliard, Phil. 2020. "HFC Files." Harvard Dataverse. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/SOK6TV>
- Himmelstein, Mark, Pavel Atanasov, and David V. Budescu. 2021. "Forecasting Forecaster Accuracy: Contributions of Past Performance and Individual Differences." *Judgment and Decision Making* 16(2): 323–62.
- Hossain, K. S. M. Tozammel, Hrayr Harutyunyan, Yue Ning, Brendan Kennedy, Naren Ramakrishnan, and Aram Galstyan. 2022. "Identifying Geopolitical Event Precursors Using Attention-Based LSTMs." *Frontiers in Artificial Intelligence* 5: 893875.
- Hossain, KT, Gao, S, Kennedy, B, Galstyan, A, Natarajan, P. 2020. "Forecasting violent events in the Middle East and North Africa using the Hidden Markov Model and regularized autoregressive models." *The Journal of Defense Modeling and Simulation*. 17(3): 269–283. <https://doi.org/10.1177/1548512918814698>
- Huang, Yuzhong, Andres Abeliuk, Fred Morstatter, Pavel Atanasov, and Aram Galstyan. 2020. "Anchor attention for hybrid crowd forecasts aggregation." In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1869–1871.
- Hyndman, Rob J, and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. Heathmont, Victoria: OTexts.
- Hyndman, Rob J., and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 27: 1–22.
- IARPA. 2011. Aggregative Contingent Estimation (ACE) <https://www.iarpa.gov/research-programs/ace>
- IARPA. 2017. Hybrid Forecasting Competition. <https://www.iarpa.gov/research-programs/hfc>
- Janssen, Christian P., Stella F. Donker, Duncan P. Brumby, and Andrew L. Kun. 2019. "History and Future of Human-Automation Interaction." *International Journal of Human-Computer Studies* 131: 99–107.
- Jin, Woojeong, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2021. "ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4636–50. Association for Computational Linguistics.
- Joseph, Regina, and Pavel Atanasov. 2019. "Predictive Training and Accuracy: Self-Selection and Causal Factors." In *Proceedings of Collective Intelligence*, 1.
- Kamar, Ece. 2016. "Hybrid Workplaces of the Future." *XRDS: Crossroads, The ACM Magazine for Students* 23(2): 22–5.
- Leigh, Andrew, and Justin Wolfers. 2006. "Competing Approaches to Forecasting Elections: Economic Models, Opinion Polling and

- Prediction Markets\*." *Economic Record* 82(258): 325–40. <https://doi.org/10.1111/j.1475-4932.2006.00343.x>
- Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2018. "Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward." *PLOS ONE* 13(3): e0194889.
- Meehl, Paul E. 1954. *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, 149. Minneapolis: University of Minnesota Press.
- Mellers, Barbara, Lyle Ungar, Jonathan Baron, Jaime Ramos, Burcu Gurcay, Katrina Fincher, Sydney E Scott, et al. 2014. "Psychological Strategies for Winning a Geopolitical Forecasting Tournament." *Psychological Science* 25(5): 1106–15.
- Merkle, Edgar C., and Robert Hartman. 2018. "Weighted Brier Score Decompositions for Topically Heterogenous Forecasting Tournaments." *Judgment and Decision Making* 13(2): 185–201.
- Montero-Manso, Pablo, George Athanasopoulos, Rob J Hyndman, and Thiyanaga S Talagala. 2020. "Fforma: Feature-Based Forecast Model Averaging." *International Journal of Forecasting* 36(1): 86–92.
- Morstatter, Fred. 2021. SAGE/HFC Geopolitical Forecasting. RCT-B. Harvard Dataverse. <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ROTHFT>
- Morstatter, Fred, Aram Galstyan, Gleb Satyukov, Daniel Benjamin, Andres Abeliuk, Mehrnoosh Mirtaheeri, KSM Tozammel Hossain, et al. 2019. "Sage: A Hybrid Geopolitical Event Forecasting System." In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 7, 6557–9. International Joint Conferences on Artificial Intelligence Organization.
- Moss, Aaron. 2022. How CloudResearch and IARPA completed the largest longitudinal online research project ever. <https://www.cloudresearch.com/resources/blog/the-largest-longitudinal-online-research-project/>
- Önköl, Dilek, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. "The Relative Influence of Advice from Human Experts and Statistical Methods on Forecast Adjustments." *Journal of Behavioral Decision Making* 22(4): 390–409. <https://doi.org/10.1002/bdm.637>
- Organisation for Economic Co-operation and Development. June 2022. OECD data. <http://data.oecd.org/united-states.htm>
- Organization of the Petroleum Exporting Countries. 2022. OPEC: Monthly Oil Market Report. June. [https://www.opec.org/opec\\_web/en/publications/338.htm](https://www.opec.org/opec_web/en/publications/338.htm)
- Parmezan, Antonio Rafael Sabino, Vinicius M. A. Souza, and Gustavo E. A. P. A. Batista. 2019. "Evaluation of Statistical and Machine Learning Models for Time Series Prediction: Identifying the State-of-the-Art and the Best Conditions for the Use of Each Model." *Information Sciences* 484: 302–37.
- Pilster, Ulrich, and Tobias Böhmelt. 2014. "Predicting the Duration of the Syrian Insurgency." *Research & Politics* 1(2): 2053168014544586.
- Prahl, Andrew, and Lyn Van Swol. 2017. "Understanding Algorithm Aversion: When is Advice from Automation Discounted?" *Journal of Forecasting* 36(6): 691–702. <https://doi.org/10.1002/for.2464>
- Rafner, Janet, Dominik Dellerman, Arthur Hjorth, Dora Veraszto, Constance Kampf, Wendy MacKay, and Jacob Sherson. 2022. "Deskilling, Upskilling, and Reskilling: A Case for Hybrid Intelligence." *Morals & Machines* 1(2): 24–39.
- Rafner, Janet, Miroslav Gajdacz, Gitte Kragh, Arthur Hjorth, Anna Gander, Blanka Palfi, Aleks Berditchevskaia, et al. 2021. Revisiting citizen science through the lens of hybrid intelligence. <https://arxiv.org/abs/2104.14961>
- Rafner, Janet, Miroslav Gajdacz, Gitte Kragh, Arthur Hjorth, Anna Gander, Blanka Palfi, Aleksandra Berditchevskaia, et al. 2022. "Mapping Citizen Science Through the Lens of Human-Centered AI." *Human Computation* 9(1): 66–95.
- Schrodt, Philip A. 2010. Automated production of high-volume, real-time political event data. APSA 2010 Annual Meeting Paper, Available at SSRN: <https://ssrn.com/abstract=1643761>
- Schutte, Sebastian. 2017. "Regions at Risk: Predicting Conflict Zones in African Insurgencies\*." *Political Science Research and Methods* 5(3): 447–65.
- Seifert, Matthias, and Allégère L. Hadida. 2013. "3 Humans + 1 Computer = Best Prediction." *Harvard Business Review*.
- Tetlock, Philip E. 2017. *Expert Political Judgment: How Good Is It? How Can We Know?—New Edition*. Princeton: Princeton University Press.
- Trouille, Laura, Chris J. Lintott, and Lucy F. Fortson. 2019. "Citizen Science Frontiers: Efficiency, Engagement, and Serendipitous Discovery with Human-Machine Systems." *Proceedings of the National Academy of Sciences* 116(6): 1902–9.
- Vasanthan, Raghavan, Aram Galstyan, Alexander G., Tartakovsky. 2013. "Hidden Markov models for the activity profile of terrorist groups." *The Annals of Applied Statistics*, 7(4): 2402–2430 December 2013. <https://doi.org/10.1214/13-AOAS682>
- Zellner, Maximilian, Ali E. Abbas, David V. Budescu, and Aram Galstyan. 2022. "A Survey of Human Judgement and Quantitative Forecasting Methods." *Royal Society Open Science* 8(2): 201187.

**How to cite this article:** Benjamin, D. M., F. Morstatter, A. E. Abbas, A. Abeliuk, P. Atanasov, S. Bennett, A. Beger, S. Birari, D. V. Budescu, M. Catasta, E. Ferrara, L. Haravitch, M. Himmelstein, K. T. Hossain, Y. Huang, W. Jin, R. Joseph, J. Leskovec, A. Matsui, M. Mirtaheeri, X. Ren, G. Satyukov, R. Sethi, A. Singh, R. Sosis, M. Steyvers, P. A. Szekely, M. D. Ward, and A. Galstyan. 2023. "Hybrid forecasting of geopolitical events." *AI Magazine* 44: 112–128. <https://doi.org/10.1002/aaai.12085>

## AUTHOR BIOGRAPHIES

**Daniel M. Benjamin** is an Assistant Professor of Decision Sciences at Nova Southeastern University, Huizenga College of Business and Entrepreneurship.

**Fred Morstatter** is a Research Assistant Professor at the University of Southern California, Information Sciences Institute.

**Ali E. Abbas** is a Professor of Industrial and Systems Engineering at the University of Southern California.





**Andres Abeliuk** is an Assistant Professor of Computer Science at the University of Chile.

**Pavel Atanasov** is a Co-Founder of Pytho LLC and an upcoming Assistant Professor of Decision Sciences at IE Business School

**Stephen Bennett** obtained his Ph.D. in Cognitive Sciences from University of California, Irvine.

**Andreas Beger** is a Political Scientist and independent Consultant working on geopolitical forecasting systems.

**Saurabh Birari** is a Machine Learning Engineer at Amazon Berlin.

**David V. Budescu** is the Anastasi Professor of Psychometrics and Quantitative Psychology at Fordham University in NYC.

**Michele Catasta** was a postdoctoral scholar at Stanford University. He is currently the Head of Applied Research at Google Labs.

**Emilio Ferrara** is a Professor of Computer Science, Communication and Preventive Medicine at the University of Southern California.

**Lucas J. Haravitch** is an Analyst at the Office of the Secretary of Defense.

**Mark Himmelstein** is a Ph.D. Candidate in Psychometrics and Quantitative Psychology at Fordham University.

**Tozammel Hossain** was a postdoctoral scholar at the USC Information Sciences Institute. He is currently an Assistant Professor in the Department of Information Science at the University of North Texas.

**Yuzhong Huang** is a Ph.D. Candidate in Computer Science at the University of Southern California, Information Sciences Institute.

**Woojeong Jin** is a Ph.D. Candidate in Computer Science at the University of Southern California.

**Regina Joseph** is a Human-Machine Interaction Researcher, Founder of Sibylink and Co-founder of Pytho LLC.

**Jure Leskovec** is a Professor of Computer Science at Stanford University.

**Akira Matsui** obtained his PhD from the USC Information Sciences Institute. He is currently an Assistant Professor of College of Business Administration at the Yokohama National University.

**Mehrnoosh Mirtaheri** is a PhD Candidate in Computer Science at the University of Southern California and Research Assistant in the Information Sciences Institute.

**Xiang Ren** is an Assistant Professor of Computer Science at the University of Southern California.

**Gleb Satyukov** is a Research Engineer at the University of Southern California, Information Sciences Institute.

**Rajiv Sethi** is a Professor of Economics at Barnard College, Columbia University and an External Professor at the Santa Fe Institute.

**Amandeep Singh** is a Senior Research Programmer at the University of Southern California, Information Sciences Institute.

**Rok Susic** is a Senior Researcher in the Computer Science Department at Stanford University.

**Mark Steyvers** is a Professor and Chancellor's Fellow at the Department of Cognitive Sciences, University of California, Irvine.

**Pedro A Szekely** is an Applied Principal Scientist at Amazon. Dr Szekely was formerly Associate Research Professor and Director of the Artificial Intelligence Division at University of Southern California, Information Sciences Institute.

**Michael D. Ward** was an Emeritus Professor of Political Science at Duke University and Founder of Predictive Heuristics; he passed away in 2021.

**Aram Galstyan** is a Research Professor of Computer Science at the University of Southern California, Information Sciences Institute.