

Task-offload Tools Improve Productivity and Performance in Geopolitical Forecasting

Ion Juvina¹, Othalia Larue¹, Colin Widmer², Subhashini Ganapathy¹, Srikanth Nadella², Brandon Minnery², Lance Ramshaw³, Emile Servan-Schreiber^{4,5}, Maurice Balick⁴, and Ralph Weischedel⁶

¹Wright State University, Dayton, OH, USA

²Kairos Research, Fairborn, OH, USA

³Raytheon BBN Technologies Corp., Cambridge, MA, USA

⁴Hypermind, LLC, Paris, France.

⁵School of Collective Intelligence, Mohammed VI Polytechnic Univ., Ben Guerir, Morocco

⁶University of Southern California, Information Sciences Institute, Los Angeles, CA, USA

ion.juvina@wright.edu, othalia.larue@wright.edu, colin@kairos-research.com, subhashini.ganapathy@wright.edu, srikanth@kairos-research.com, brad@kairos-research.com, lance.ramshaw@raytheon.com, emile@lumenogic.com, mbalick@lumenogic.com, and weisched@isi.edu

Abstract

Recent studies in geopolitical forecasting have identified psychological variables that predict forecasting accuracy. We studied the effect of providing human forecasters with automated information search and task management support tools. Our research aimed to determine whether use of the support tools could explain additional variance in forecasting performance above and beyond psychological variables. We found that the provided tools encouraged participants to do more work (i.e., information search, communication, reflection, etc.), which in turn resulted in improved forecasting performance.

Background

Forecasting and other forms of intelligence analysis are information-intensive tasks that rely heavily on information foraging and sense-making tools (Pirulli & Card, 2005). However, forecasting is more challenging than other investigational search and sense-making tasks. In a typical investigational task, the answer exists somewhere, and the users have to find their way to that answer or assemble an answer from pieces of information found in different locations. In forecasting tasks, the answers do not exist yet; they have to be constructed by the users. An element of novelty is always present in forecasting; no forecasting solution applies to more than one problem, even though general strategies may exist. Typically, real-world forecast-

ing occurs over an extended time course, during which the world changes and potentially relevant but also irrelevant or misleading evidence accumulates. Further adding to the complexity, forecasters often engage in multiple forecasting tasks, and each task may be attempted by a group of cooperating and/or competing forecasters.

The symbiosis between humans and machines (Licklider, 1960) holds great promise for tackling the unparalleled complexity of the forecasting task. The science and practice of human-technology coordination have departed from the traditional function allocation methods (who-does-what or men-are-better-at / machines-are-better-at, MABA/MABA; Fitts, 1951) and is currently moving toward a *human-technology teaming* approach in which the focus is on how machines can become effective team players (Dekker & Woods, 2002) and how humans and technology co-evolve (Ackerman, 2000).

The tools used in our study are called *hybrid* because they are intended to combine human and machine capabilities (Rahwan, Cebrian, et al., 2019) to improve the performance of the whole socio-technical system that generates forecasts. Using hybrid tools to assist forecasting serves three purposes: (1) correct for cognitive biases; (2) reduce the cognitive load of forecasters; and (3) increase the amount of relevant information available to the forecaster. These goals can be complementary and mutually reinforcing: providing humans with machine-made forecasts and making the relevant information easier to search and interpret may reduce cognitive load and cognitive biases, which

in turn facilitates high quality forecasts, which via various aggregation methods result in better “hybrid” forecasts.

Cognitive workload and fatigue have been shown to affect judgment quality, with forecast quality decreasing as the number of forecasts made in a day increased. As they get fatigued, forecasters exhibit more herding behavior and less granularity in their forecasts (Hirshleifer et al., 2019). Task-offload tools can be used to delegate some task demands to automation (Kirlik, 1993). However, externalizing too much task-related information can reduce the user’s ability to meaningfully engage in high-level processes such as planning and reasoning and may harm motivation and performance (Van Nimwegen, Burgos, Van Oostendorp, & Schijf, 2006). Thus, a hybrid tool that aims to support human forecasting must strike a balance between offloading task demands and maintaining user engagement.

Method

Human forecasters had to solve forecasting problems (FPs) about real-world events in the following domains: conflict, economics, health, politics, science, and technology. Participants were asked to provide an initial forecast and update it as many times as necessary based on information they searched for, updates of this information, or new information. Each forecasting problem had between two and five discrete, mutually exclusive outcome options. Outcome options had to be assigned a probabilistic forecast with probabilities over all options adding up to 1.

Two samples of participants were recruited for this study. The first sample consisted of volunteers with interest in geopolitical analysis. They were mostly U.S. citizens (76%), males (82%), with an average age of 43, and with a relatively high level of education (53% had received a postgraduate degree). A second sample of participants was recruited from the members of the web service TurkPrime, typically referred to as workers. To simplify our language, we will refer to the participants from the first sample as *Volunteers* and to the participants from the second sample as *Turkers*.

Forecasting performance was measured with the Brier score (Brier, 1950) and a relative accuracy score. The Brier score provides a measure of the error of a probability forecast: the further a forecast probability is from the actual outcome, the larger the error:

$$\text{Brier score} = \sum (p_i - o_i)^2$$

Where p_i is the probability assigned to answer i , and o_i is 1 if answer i is correct, or 0 if it is not. The Brier score is between 0 (perfect forecast) and 2 (worst possible forecast).

The accuracy score is a relative score based on one’s Brier scores compared to the median Brier scores of all

participants. The accuracy on a particular day is $c_a - y_a$, where y_a is the participant’s Brier score on that day, and c_a is the crowd’s *median* Brier score on that day. The accuracy score varies between -2 (worst) and 2 (best).

The participants accessed a dedicated website containing hybrid features designed to assist them with information search and task management¹. The use of the available features was optional to users. We assumed users would strategically (Kirlik, 1993) choose the features they needed depending on what stage of the task they needed more support with (Huurdean, Kamps, & Wilson, 2019) or what costs and benefits they attributed to using automated tools (Pirolli & Card, 2005). Only a subset of these features was used in the study that we report here. The participants could access numerical indicators relevant to the selected FP, other user forecasts, forum conversations, news, links, tabs, and so on (see Table 1).

The Indicators tool displays a list of indicators, which are statistics relevant to the FP. Indicators can be economic statistics, Internet search term frequency, information from databases, etc. A participant can monitor how their indicators change over time to see when something changes about a FP and decide to update their forecast.

The Query tool allows the participants to extract data from several relevant sources. Query bots automatically access web sites and databases providing the current and past trends of indicators underlying many of the FPs. To guide participants to queries that would help them answer a given FP, the system automatically recommends databases to participants. A set of databases was pre-compiled by subject matter experts for each FP category (e.g. conflict, economy, or health); when a FP from a certain category is posted, the system automatically recommends the databases for that category. Participants can edit a suggested query, for example, by modifying some of the suggested values. Participants can also manually add databases they deem relevant to a given FP. Then they can create queries on databases using a query editor that allows them to specify a date, location, type, actor, etc.

Forecasters can save a query in order to automatically track its results in time. A saved query becomes an Indicator. Every six hours, the system automatically reruns the query. Forecasters can also manually rerun their queries. Indicators can be shared among forecasters by making them public. Indicators updated over the course of an FP’s lifecycle are viewable to participants as time-series graphs.

Another important feature allows participants to create custom alarms (also called alerts) based on indicators. Alarms can alert a participant when key statistics have updated that may affect his or her forecast. Alarms are created with the Alarm Rule Editor. They are written in the form of IF condition, THEN action. That is, the participant

¹ The original experiment included a control group that did not have access to these hybrid features. However, the data from the control group were not available at the time this paper was written.

specifies the conditions that trigger the alarm and what actions should be taken once the alarm is triggered (i.e., forecast recommendations). The participants can create three types of alarms: crowd-based, indicator-based, and time-based alarms. Crowd-based alarms track the average forecast among all forecasters for a specific outcome and will alert the participant when the crowd’s prediction has changed. Indicator-based alarms track the value of one or more indicators. Once an indicator reaches a pre-specified value, the participant is notified. Time-based alarms remind the user to review their forecast after a specified period has passed. Email updates were sent to the participants when their alarms fired.

Table 1. Hybrid tools supporting forecasting

Feature	Purpose / Function
Question Details	Provide additional information about the question.
Crowd Forecast History	Provide aggregate information about how all forecasters have answered the question.
Indicators	Display current value and time course of statistics relevant to the FP.
Forum	Allow participants to discuss the question and share information
Links	Display a list of useful links to sources relevant to the question.
News	Suggest relevant news and allow news search.
Query	Allow participants to extract data from relevant sources. Query bots automatically recommend relevant data sources and queries. Query editor supports creation and reruns.
Customized Alarm	Notify participant when relevant information (e.g., the value of a particular indicator) changes and recommend a forecast update.
Customized Rule of Thumb	Detect change in relevant information, automatically update forecast, and notify the participant.
Email Updates	Provide participants with general information and customized recommendations and feedback about their forecasts.

Most of the participants completed the Cognitive Reflection Test (Frederick, 2005), the Actively Open-minded Thinking scale (Stanovich and West, 1997) and the Need for Cognition scale (Cacioppo et al., 1984). These variables were found in previous studies to correlate with forecasting performance (Mellers et al., 2015). In addition, we collected extensive data on participant behavior, such as the number of FPs forecasted, the number of forecast up-

dates per FP, frequency of usage for each hybrid feature, etc.

We expected that the provided suite of hybrid features would improve forecasting productivity and quality, that is, the number of forecasts participants can generate, the frequency at which these forecasts can be updated, and the accuracy of these forecasts. The hybrid features should allow participants to reduce the cognitive load associated with monitoring their forecasts and updates, which in turn should allow them to make more forecasts and focus on evaluating information quality and relevance. For example, when alarms trigger, they remind participants to update their forecasts, and a higher frequency of updating has in turn been linked to better forecasting performance (Tetlock & Gardner, 2015). When users create alarms, they are implicitly encouraged to employ a top-down (model-driven) strategy. They need to develop intuitive causal models of what factors determine the occurrence of the event to be forecasted. Due to the nature of the forecasting task, modeling and understanding the (hidden) causes of events are critical for performance.

Results

To evaluate if the use of hybrid features improved forecasting productivity and performance, we split the forecasters into two groups: one that used no hybrid features (queries, indicators, or alarms) composed of 519 participants and a group of participants who used one or more hybrid features, 319 participants.

The average number of forecasts per FP was higher for participants using the hybrid tools, $t(371.67) = -6.44, p < 0.001$. Thus, participants who used hybrid tools made more forecast updates. The average number of FP topics forecasted was also higher for participants who used hybrid tools, $t(737.78) = -8.81, p < 0.001$. Thus, the participants who used hybrid tools attempted to forecast a wider range of IFP topics. The total number of forecasts submitted was higher for participants who used hybrid tools, $t(328.61) = -4.89, p < 0.001$. Forecasting performance as measured by the Brier score and the relative accuracy measure (described above) was higher for the participants who used hybrid tools, $t(835.29) = 1.99, p = 0.05$ for Brier scores and $t(834.07) = -4.58, p < 0.001$ for relative accuracy.

Thus, as expected, forecasting productivity and accuracy were higher in those participants who used the provided hybrid features. However, it remains unclear whether these findings are driven by the availability of hybrid features or by motivation. Mellers et al. (2015) found that the frequency of forecast updating, which they considered to be a behavioral indicator of motivation, was a significant predictor of forecasting performance. Arguably, the direction of causality could go both ways: (1) the highly motivated participants made a larger number of forecast updates and used the provided hybrid tools, which in turn increased perfor-

mance, or (2) the hybrid tools increased the participants motivation to make updates, which in turn increased performance.

To test these two possibilities, we constructed and tested two structural equation modeling (SEM) models attempting to explain the structural relations between hybrid tools usage (a sum of queries, indicators, and alarms used), psychometric measures (cognitive reflection – cRS and actively open-minded thinking – aTS), motivation (number of topics forecasted – N and average number of forecasts per IFP - aFI) and performance (Brier score – Brr and accuracy – Acc).

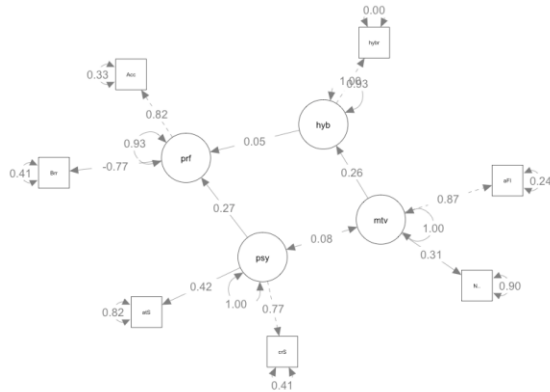


Figure 1. SEM Model 1: Motivation causes hybrid feature use, which improves performance. Circular nodes indicate latent variables. Square nodes indicate manifest variables. Unidirectional arrows indicate regressions while bidirectional arrows indicate correlations. All paths were significant except the path from hybrid feature usage to performance (hyb->prf). See appendix 1 for a higher-resolution diagram.

Model 1 (Fig. 1) hypothesizes a direct causal link between hybrid feature use and forecasting performance, whereas model 2 (Fig. 2) hypothesizes an indirect causal link (via motivation) between hybrid feature use and forecasting performance. Model 1 assumes that motivation causes hybrid tool usage, which in turn causes increased performance. It also includes the known associations between psychometrics, motivation, and forecasting performance. Model 2 assumes that hybrid tool usage causes motivation, which in turn causes increased performance. Similar to model 1, model 2 also includes the known associations between psychometrics, motivation, and forecasting performance.

We compared the two models using the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Model 2 had AIC = 15913 and BIC = 15994, whereas Model 1 had AIC = 15941 and BIC = 16022, thus Model 2 fits the data slightly better than Model 1.

Model 2 supports the hypothesis that the use of hybrid tools has a direct effect on motivation. Conceivably, email alerts about indicator changes and crowd changes motivated participants to update their own forecasts and perhaps

do additional information searches. In agreement with previous studies, motivation had a direct effect on performance, as did the psychometric variables actively open-minded thinking and the tendency to engage in cognitive reflection.

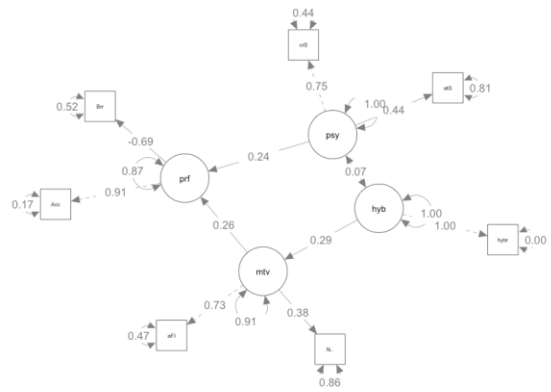


Figure 2. SEM Model 2: Hybrid feature use causes motivation, which causes performance. Circular nodes indicate latent variables. Square nodes indicate manifest variables. Unidirectional arrows indicate regressions while bidirectional arrows indicate correlations. All paths were significant. See appendix 2 for a higher-resolution diagram.

Discussion and Conclusion

Previous studies (Mellers et al., 2015; Tetlock & Gardner, 2015) reported dispositional and behavioral predictors of forecasting performance. These findings were replicated in our study: cognitively reflective and open-minded participants made better forecasts. In addition, Mellers et al. (2015) showed that participants who updated their forecasts more often achieved better forecasting performance. This finding was also replicated in our study.

Our study added a suite of hybrid feature to assist forecasters with the laborious tasks of information search, sense making, and decision-making. The use of these tools was optional. We assumed users would act strategically (Kirlik, 1993) and use these tools as needed. The expectation was that forecasters equipped with hybrid tools would become more productive and more accurate. The effect of the hybrid tools was expected to be independent of the effects that were already known (i.e., cognitive ability, cognitive style, and motivation). For example, hybrid tools were expected to be helpful above and beyond a participant’s motivation or cognitive ability. What we found does not entirely support this expectation. We did find that the use hybrid features improve forecasting performance, but this relationship is most likely mediated by motivation. The use of hybrid features increased the forecasters’ productivity, as indicated by the number and the variety of IFPs they forecasted and the frequency of forecast updates. Since the use of hybrid features was optional, the relation-

ship between the use of hybrid features and forecasting performance must be interpreted with caution, as only a minority of participants used the provided hybrid tools (319 of 839) and the decision to use hybrid features might be confounded by other factors such as trust in automation and in other forecasters (Juvina, Collins et al., in press).

Our SEM analysis provided support for the interpretation that the provided hybrid features encouraged the participants to do more work (i.e., information search, communication, reflection, etc.), which in turn resulted in improved forecasting performance.

We focused here on a subset of hybrid tools, namely queries, indicators, and alarms. They appear to be useful in driving improvements in forecasting performance. While it is not surprising that supporting users information foraging and sense making improves forecasting performance, our unique contribution emphasizes the importance of engaging users in creating their own support tools. We provided the alarm editor to encourage participants to create customized alarms that would alert them when potentially relevant information changes and recommend a forecast update. The participants who chose to create an alarm had to specify the conditions that would trigger the alarm (i.e., specific changes in one or more indicators) and the action to be recommended (i.e., a specific change in the forecast). Arguably, the alarm editor challenged participants to create their own intuitive models of information search and forecasting and turn these models into support tools. The results highlight the importance of providing tools that are not only useful and useable, but are able to engage users and enhance their cognitive activity, aiming to strike a balance between user effort and information search automation (Bates, 1990), ultimately achieving the goals of human machine symbiosis and co-evolution (Licklider, 1960; Ackerman, 2000).

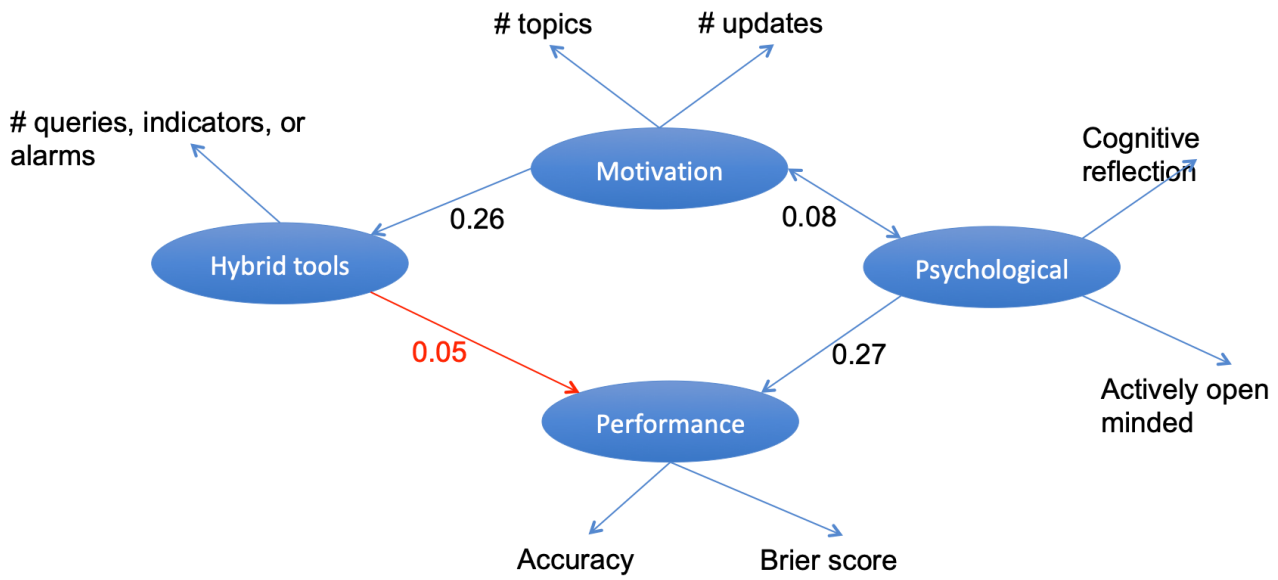
Acknowledgements

This research was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract no. 2017-17072100002 to Raytheon-BBN, through subaward to Kairos Research. 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 or BBN. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Ackerman, M.S. (2000). The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2): 179-203.

- Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing & Management*, 26(5):575-591.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3.
- Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of personality assessment*, 48(3), 306-307.
- Dekker, S.W.A & Woods, D.D. (2002). MABA-MABA or Abracadabra? Progress on Human-Automation Co-ordination. *Cognition, Technology & Work* 4:240-244.
- Fitts, P.M. (ed.) (1951). Human engineering for an effective air navigation and traffic control system. National Research Council, Washington, DC.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4), 25-42.
- Hirshleifer, D., Levi, Y., Lourie, B., & Teoh, S. H. (2019). Decision fatigue and heuristic analyst forecasts. *Journal of Financial Economics*.
- Huurdeman, H. C., Kamps, J. & Wilson, M. L. (2019). The multi-stage experience: the simulated work task approach to studying information seeking task stages. In *Proc. BIIRRR workshop at CHIIR 2019*.
- Juvina, I., Collins, M.G., Larue, O., Kennedy, W., de Visser, E., & de Melo, C. (in press). Toward a unified theory of learned trust in interpersonal and human-machine interactions. *ACM Transactions in Interactive Intelligent Systems*.
- Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human factors*, 35(2), 221-242.
- Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, volume HFE-1, pages 4-11.
- Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., ... and Tetlock, P. (2015a). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of experimental psychology: applied*, 21(1), 1.
- Pirolli, P. & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Intelligence Analysis*, 2-4.
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochele, H., Lazer, D., Mcelreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behavior. *Nature*, 568, 477-486.
- Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89(2), 342.
- Tetlock, P.E. & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. New York: Broadway Books.
- Van Nimwegen, C., Burgos, D., Van Oostendorp, H., & Schrijf, H. (2006). The paradox of the assisted user: guidance can be counterproductive. In *Proc. CHI 2006*, 917-926.
- Appendix 1: Higher-resolution diagram for SEM model 1.



Appendix 2: Higher-resolution diagram for SEM model 2.

