Future Indices

How Crowd Forecasting Can Inform the Big Picture

CSET Issue Brief

AUTHORS
Michael Page
Catherine Aiken
Dewey Murdick

Executive Summary

What will the world look like in 2025? How will changing geopolitical and tech-security trends—such as U.S.-China relations, remote work, and public interest in automation—shape the world we occupy? These questions about tomorrow are on the minds of policymakers today. Presenting recent forecast data collected through CSET's Foretell project (cset-foretell.com), this brief shows how crowd forecasting can inform policy by providing data on future trends and linking those trends to future policy-relevant scenarios.

We illustrate Foretell's methodology with a concrete example: First, we describe three possible *scenarios*, or ways in which the tech-security landscape might develop over the next five years. Each scenario reflects different ways in which U.S.-China tensions and the fortunes of the artificial intelligence industry might develop. Then, we break each scenario down into near-term *predictors* and identify one or more *metrics* for each predictor. We then ask the crowd to forecast the metrics. Lastly, we compare the crowd's forecasts with projections based on historical data to identify *trend departures*: the extent to which the metrics are expected to depart from their historical trajectories.

> **Key Terms**
>
> **Scenario.** A qualitative description of the policy environment three to seven years from now.
>
> **Predictor.** A driver of a scenario that might manifest in the next year.
>
> **Metric.** An observable measure of a predictor. Multiple metrics are typically used to measure a predictor.
>
> **Trend Departure.** The difference between where the crowd forecasts a metric will go and a projection based on historical data.

Our preliminary findings suggest two outcomes—both involving increasing U.S.-China tensions and Department of Defense AI R&D investments—are most likely. Forthcoming data on commercial AI R&D investments, globalization, and industry-DoD tensions will inform which of these two scenarios is more likely.

Foretell's approach is a variation on a proposal by Philip E. Tetlock, co-founder of the Good Judgment Project, which won the Intelligence Advanced Research Projects Activity (IARPA)-funded Aggregative Contingency Estimation (ACE) forecasting tournament. We believe a scaled-up version of Foretell would contribute to a more evidence-based policymaking environment.

Table of Contents

Introduction

Experts disagree about what the world will look like in five years. In terms of geopolitical competition, some experts predict an ascendant China,[1] others predict the collapse of the Chinese Communist Party,[2] and others predict a U.S.-China cold war.[3] While expert predictions are indispensable, they are limited. It is difficult to build on them, locate where others might disagree, and identify what evidence supports or undermines them. Yet such predictions are critical tools to help frame policy debates. Therefore, greater insight on the relative likelihoods of qualitative expert predictions is of great value to policymakers.

To address this gap, CSET launched Foretell, a pilot project that uses crowd forecasting and data analytics to inform tech-security policy.[4] Our aim is to build on big picture scenarios, such as qualitative expert predictions, by making them more amenable to quantitative analysis. Building on previous research, notably IARPA's ACE forecasting tournament, Foretell relies on the *wisdom of the crowd*—the collective opinion of a large group—to generate probabilistic forecasts on specific, near-term questions.[5] But as noted by the founders of the winning ACE team, Philip E. Tetlock and Barbara Mellers, along with J. Peter Scoblic,"the specificity required to make questions rigorously resolvable precludes asking 'big' questions."[6]

Tetlock, Mellers, and Scoblic were describing an example of the problem of measurement: the gap between what we *can* measure and what we *want* to measure. To generate quantitative insights into big-picture concepts such as "U.S.-China tensions" or "a strong tech sector," we must identify observable *metrics* that approximate the concept.

To address both the prediction and measurement problems, Tetlock proposed a series of tournaments to generate "clusters of short-term questions that, taken individually, are rigorously resolvable but that can collectively tip the scales of plausibility in high-stakes debates."[7]

We are implementing a simplified version of this method on Foretell by selecting metrics that inform big picture scenarios and aggregating the extent to which, for each metric, the crowd forecasts depart from their historical trajectories (*trend departure*). Section I discusses an example application, linking three possible scenarios to a set of metrics for which we have forecast

data. Section II overviews the methodology in more detail. Section III discusses where we go from here.

What Will the Tech-Security Landscape Look Like in 2025?

To launch Foretell, we developed three scenarios depicting what the tech-security landscape might look like in 2025.[8] We focused on two important issues for tech-security policy, U.S.-China tensions and the fortunes of the U.S. artificial intelligence industry. Table 1 shows how by varying these issues, we could end up in very different worlds.

Table 1. Three Possible 2025 Scenarios

	U.S.-China tensions increase	U.S.-China tensions decrease
AI industry booms	Tense Economic-Security Balance	Virtually Integrated
AI industry declines	Domestic & Securitized	*Omitted*

In the first scenario (*Tense Economic-Security Balance*), U.S.-China tensions have risen as the AI industry has grown, leading to recurring conflicts between economic and security goals and deep divisions between the Department of Defense and the AI industry. In the second scenario (*Virtually Integrated*), U.S.-China tensions have subsided as the AI industry has flourished, aided by the public's heightened interest in automation and the development of remote, global workforces. In the third scenario (*Domestic & Securitized*), U.S.-China tensions have risen as the economy has deteriorated, leading to a less globalized economy and an AI industry more reliant on defense funding. We did not include a scenario in which tensions decrease and the AI industry declines because we determined, after consultation with experts, that this is an unlikely scenario.

Each of these scenarios calls for a different policy approach. For example, a virtually integrated AI industry impacts the role of export controls, and a remote AI workforce impacts immigration policy. Because a variation on each scenario is possible, the best policy approaches are robust to all scenarios,

but robustness comes at a cost. The more probable one of these scenarios becomes relative to the others, the more policy should target that scenario, and the less important it is to hedge on the others. The key question then is how likely the different scenarios are, and how we will know when relative likelihoods change.

It's difficult to assess the relative likelihoods of our 2025 scenarios for two reasons. First, they describe events three to seven years from now, and forecasting is most accurate over shorter time periods. And second, they are complex and not directly observable. Whether analyzing historical events or future events, quantitative methods require well-defined observables.

To solve these problems, we broke each scenario down into predictors and metrics, and posed the metrics to the crowd as forecast questions. Section II describes this process in more detail (see Figures 1 and 2). We then used the crowd forecasts to identify trend departures, meaning areas in which the policy environment appears to be changing faster or slower than one would expect based on projections from historical data. Table 2 shows trend departures for our 2025 scenarios based on a sampling of the metrics and predictors identified for each scenario.

Table 2: How Foretell aggregates crowd forecasts to inform big picture scenarios

Metric [trend departure]		Predictor [trend departure]		Scenario		
				Domestic & Securitized	Virtually Integrated	Tense Economic-Security Balance
Decreasing U.S.-China trade[9]	0.9	Increasing U.S.-China tensions	0.6	✅ Increasing U.S.-China tensions are a predictor	❌ Decreasing U.S.-China tensions are a predictor	✅ Increasing U.S.-China tensions are a predictor
Decreasing Chinese O visas[10]	1.1					
Increasing unfavorable public view on China[11]	0.5					
Increasing Chinese incursions of Japanese air space[12]	0					
Increasing DOD AI R&D contracts[13]	0.4	Increasing DOD AI R&D investment	0.2	✅ Increasing DoD AI R&D investment is a predictor	❌ Decreasing DoD AI R&D investment is a predictor	✅ Increasing DoD AI R&D investment is a predictor
Increasing DOD AI grants[14]	0					
Increasing big tech revenue[15]	0.3	Increasing commercial AI R&D investment	0.2	❌ Decreasing commercial AI R&D investment is a predictor	✅ Increasing commercial AI R&D investment is a predictor	✅ Increasing commercial AI R&D investment is a predictor
Increasing private tech fundraising[16]	0.3					
Increasing machine learning job postings[17]	0					
Increasing big tech H-1B visas[18]	-0.3	Increasing skilled-labor migration	-0.3	✅ Decreasing skilled-labor migration is a predictor	❌ Increasing skilled-labor migration is a predictor	Not a predictor
Increasing remote software engineer jobs[19]	5.2	Increasing remote tech economy	5.2	❌ Stable or decreasing remote tech economy is a predictor	✅ Increasing remote tech economy is a predictor	Not a predictor

Source: Foretell. For the underlying data and model, see the Foretell GitHub repository, https://github.com/georgetown-cset/public-foretell.

While these preliminary results are best understood in combination with more conventional analytical tools, the results are illustrative. They currently point to a close battle between the *Domestic & Securitized* scenario and *Tense Economic-Security Balance* scenarios, as reflected by the green check marks in Table 2. The crowd to-date expects U.S.-China tensions and DOD AI R&D investment to increase relative to historical trend projections, both of which are **predictors** of these scenarios and indicate movement away from the *Virtually Integrated* scenario. Meanwhile, a significant differentiator of the *Domestic & Securitized* and *Virtually Integrated* scenarios is commercial AI R&D investment, for which we presently see mild trend departure favoring *Tense Economic-Security Balance*. Forthcoming forecast questions on globalization and DoD-industry tensions will provide predictors that help identify which of these two scenarios is more likely.

Our Methodology

Below is the five-step process we are using to aggregate historical and forecast data to inform big picture scenarios, as illustrated in Table 2. Table 3 summarizes the steps. The data and model underlying Table 2 are available on GitHub.[20]

Table 3: Collecting and aggregating crowd forecast data

| Step 1: Decompose scenarios into predictors	Step 2: Identify metrics for the predictors	Step 3: Collect historical and forecast data	Step 4: Estimate trend departure	Step 5: Aggregate
Break down a policy-relevant scenario into the trends that precede it	Find one or more metrics that adequately capture each predictor	For each metric, collect historical data points and ask the crowd to forecast future data points	Compare crowd forecasts with projections from historical data; look for divergence	Aggregate trend departure across metrics to inform the likelihood of policy-relevant outcomes

Step 1: Decompose scenarios into predictors

We begin with a big picture *scenario*: a description of the policy environment approximately three to seven years from now. Scenarios can be constructed in multiple ways. In the example in Section I, we used the 2x2 matrix

technique. An alternative approach is to begin with qualitative expert predictions. For example, experts regularly make predictions about the future of U.S.-China relations,[21] or whether we're heading toward a high--tech dystopia.[22] Such expert predictions can serve as starting points for the quantitative analysis described here.

We then break the scenarios down into *predictors*: the near-term drivers of the scenarios of interest. Figure 1 shows the predictors for the *Domestic & Securitized* scenario from Section I.

Figure 1. Breaking scenarios down into predictors
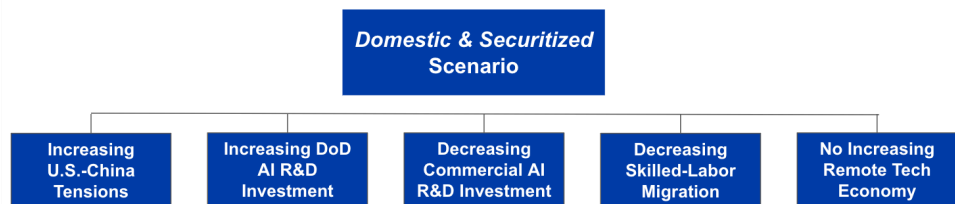
Step 2: Identify metrics for predictors

Because most predictors are not directly observable, we identify metrics that, alone or in combination with others, approximate the predictor. For example, quantity of AI publications is a common metric for quantity of AI research.[23]

For more complex predictors, multiple metrics can form an index that approximates the concept of interest. For example, as shown in Figure 2, the predictor of increasing U.S.-China tensions could be measured by trade levels, immigration flows, public opinion, and military actions.[24]

Figure 2. Identifying metrics for predictors
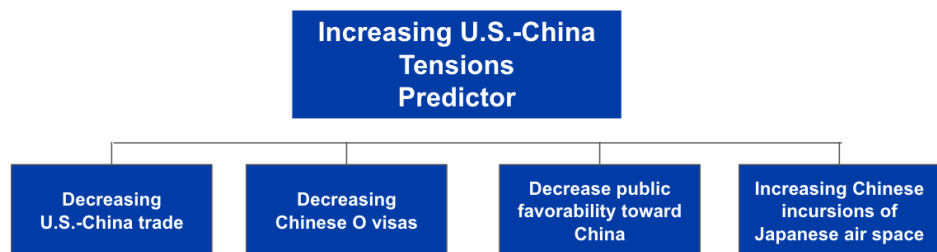
*Step 3: Collect historical and forecast data*

The next step is to gather data for the metrics. We first collect data on the metrics' historical values. By projecting the historical values forward, we create a baseline for the trend departure measure discussed in Step 4.

We then provide the historical values to the crowd and ask them to forecast the metrics' future values. Foretell currently has about 1,000 registered forecasters, comprising primarily graduate students in relevant fields.[25] Figure 3 provides an example of the historical data we collect and make available to the forecasters, supplemented with real-time forecast data.

Figure 3: The data we collect and provide to forecasters

U.S. Job Postings Requiring Machine Learning Skills
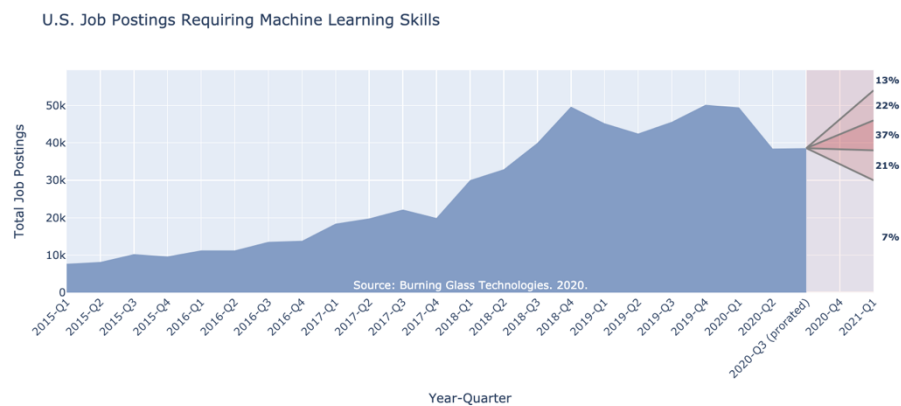
Step 4: Estimate trend departure

At this point, we have two forecasts: one based entirely on historical data (*historical projection*) and the other from the crowd (*crowd forecast*). The difference between the two is the trend departure.

Trend departure can be understood as a surprise factor, a signal of whether a metric's value should cause an analyst to stop and reconsider their assumptions. Consider the U.S.-China trade metric in Figure 4. What's noteworthy about 2019 is not its absolute value, $560 billion, or even that the 2019 value is 15 percent lower than the 2018 value. Rather, what's noteworthy is that the 15 percent decrease in 2019 deviated so significantly from historical trends, coming in $109 billion below the historical projection.

Figure 4. U.S.-China trade in 2019 was $109 billion below the historical projection
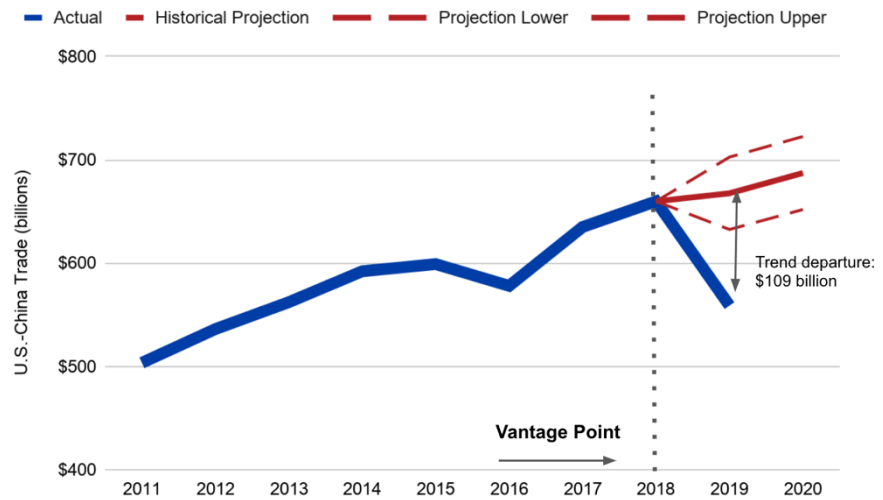
The actual (solid blue) is U.S. Census Bureau data. The projection (solid red) is based on the AAA ETS (exponential smoothing) algorithm. The upper and lower bounds (dashed red) are that projection's 95 percent confidence interval.

The example in Figure 4 involves a historical data point, U.S.-China trade in 2019, but trend departure can be calculated in the same manner for forecasted data points. In the case of U.S.-China trade, as shown in Figure 5, the crowd forecasts a 2020 value of $491 billion and a 2021 value of $505 billion, $131 billion and $142 billion below the historical projections, respectively.

Figure 5. The crowd forecasts significant trend departure in U.S.-China trade in 2021
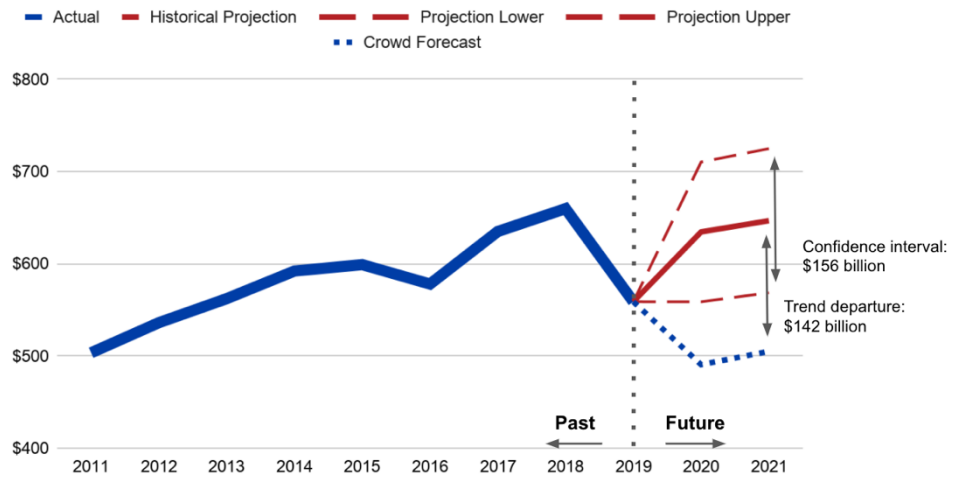
The actual (solid blue) is U.S. Census Bureau data, and the projection (dashed blue) is Foretell forecast data. The historical projection (solid red) is based on the AAA ETS (exponential smoothing) algorithm. The upper and lower bounds (dashed red) are that projection's 95 percent confidence interval.

Step 5: Aggregate

Finally, we put all the pieces back together. To create a common scale across metrics with different levels of variation, we divide trend departure by the historical projection's confidence interval. The confidence interval provides a measure of what range of values is expected in light of a metric's historical values. If the historical trend is very consistent, the confidence interval will be small and therefore moderate departures surprising. That's the case in Figure 4, which shows a $71 billion confidence interval. By contrast, if the historical values vary greatly, the confidence interval will be large and therefore moderate departures less surprising. That's the case in Figure 5, which shows a larger $156 billion confidence interval for 2021 after taking into account the anomalous 2019 value. Therefore, although the absolute trend departure amount is greater in Figure 5 (2021 forecast) than in Figure 4 (2019 actual), after dividing by the confidence interval, the trend departure is greater in 2019 (1.5) than in 2021 (0.9).

We can aggregate trend departure at the metric or predictor level. Table 1 shows aggregation at the metric level. For the U.S.-China Tensions predictor,

for example, the crowd expects that each of the four metrics will increase relative to their historical trends, forecasting trend departures between 0 and 1.1. Taking the simple average of the four metrics yields a predictor-level trend departure of 0.6. Alternatively, we could give the metrics different weights. For example, if trade seems particularly important and Chinese incursions of Japanese airspace unimportant, we could give the former a weight of 50 percent, the latter a weight of 10 percent, and the others weights of 20 percent, which would yield a predictor-level trend departure of 0.8.

Aggregating trend departure at the predictor level works in the same manner, meaning, in principle, we could quantify the extent to which the crowd thinks we are heading toward one scenario or another. Until we have more metrics and predictors to capture the scenario of interest, however, adding a second level of quantification magnifies sources of error—such as the selection and weighting of predictors and metrics—without adding offsetting insight.

Sensitivity to Technical Choices

This brief describes our methodology at a conceptual level, but to implement it, we made many technical choices, such as what algorithm to use to create the historical projection, how to calculate trend departure, and how to standardize trend departure values across metrics with different levels of historical variation. Our underlying data and the model used to generate our results is publicly available and we encourage others to improve upon our technical choices.

Ultimately, however, many of these technical choices are incidental to the results. A virtue of our focus on trend departure is that we are interested only in big changes, and the big changes should not be sensitive to debatable technical choices.

Looking Forward

We believe using crowd forecasting to inform big picture scenarios can improve policymaking in two ways. First, it can foster productive disagreement by helping policy analysts identify where they disagree and what data would advance the debate. Policy disagreements might be more manageable once reduced to specific, measurable uncertainties, such as the effect an export control would have on U.S. semiconductor manufacturers.

Second, it could serve as a warning system. In a complex, dynamic environment, it can be difficult to appreciate the significance of ongoing change. For example, two policymakers might disagree about whether we are heading toward the *Domestic & Securitized* scenario described in Section I, but agree about what to do if we are headed toward that scenario. By continually monitoring and forecasting dozens of metrics that inform the likelihood of the *Domestic & Securitized* outcome, we can effectively automate our ability to notice changes that should trigger a reconsideration of strategic policy.

Many potential obstacles remain. Among others, this methodology assumes forecasters are acting in good faith. If instead, forecasters attempt to manipulate the results to achieve their desired policy goals, they would undermine the integrity of the system. We believe such risks are best addressed in concrete cases, and in the abstract, do not pose enough risk to offset the potential upside.

Foretell is still at the proof-of-concept stage. Over the remainder of its pilot, we will study the crowd's accuracy relative to subject-matter experts and see what insights we can extract from other crowd-generated data, including rationales and the full distribution of probabilities, rather than simply the point estimates. However, for this method to realize its ultimate potential, scale is necessary. With sufficient policymaker interest, we believe a scaled-up forecasting project—with more metrics, forecasters, and end-use applications—will flourish.

Acknowledgments

Document Identifier: doi: 10.51593/20200093

Endnotes

[1] Michael Morell, "Analysis: The national security implications of COVID-19," CBS News, May 8, 2020, https://www.cbsnews.com/news/coronavirus-national-security-implications-analysis/.

[2] Minxin Pei, "China's Coming Upheaval," Foreign Affairs, May/June 2020, https://www.foreignaffairs.com/articles/united-states/2020-04-03/chinas-coming-upheaval.

[3] John Haltiwanger, "The US and China are on the brink of a new Cold War that could devastate the global economy," Business Insider, May 14, 2020, https://www.businessinsider.com.au/the-us-china-entering-new-cold-war-amid-coronavirus-2020-5.

[4] To learn more about Foretell, check out cset-foretell.com. Foretell's approach is a variation on a proposal by Philip Tetlock, co-founder of the Good Judgment Project, which won the IARPA-funded Aggregative Contingent Estimation (ACE) forecasting tournament.

[5] ACE was a four-year tournament designed to identify what methods are most effective at estimating the likelihood of geopolitical events between one month and one year in the future. The Good Judgment Project, a team led by University of Pennsylvania professors Phillip Tetlock and Barbara Mellers, won the tournament by recruiting thousands of volunteers, training them to forecast, and aggregating their judgments. Tetlock and Dan Gardner describe their approach in Tetlock, Gardner, *Superforecasting: The Art and Science of Prediction* (Broadway Books, 2015).

[6] Tetlock et al., "Bringing probability judgments into policy debates via forecasting tournaments," Science Vol 355 Issue 6324, (February 5, 2017), 481-483, https://science.sciencemag.org/content/355/6324/481.full. See also Scoblic & Tetlock, "A Better Crystal Ball: The Right Way to Think About the Future," Foreign Affairs, November/December 2000, https://www.foreignaffairs.com/articles/united-states/2020-10-13/better-crystal-ball.

[7] Philip E. Tetlock, "Full-Inference-Cycle Tournaments: The Quality of our Questions Matters as Much as the Accuracy of our Answers," Prepared for IARPA, August 30, 2017, available through Dropbox at https://t.co/dLO0CXac8A?amp=1. Scoblic & Tetlock, supra note 6, makes a similar proposal, suggesting how qualitative scenario planning and probabilistic forecasting can be combined to create warning systems for policymakers.

[8] Aiken, Page, "Three Possible 2025 Worlds that Should Inform Policy Today," Foretell, June 11, 2020, https://www.cset-foretell.com/blog/three_worlds.

[9] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/76-what-will-total-u-s-trade-volume-with-china-imports-and-exports-of-goods-be-in-2021. Data on a related question covering an earlier time period is also available, https://www.cset-foretell.com/questions/22-how-much-total-trade-will-there-be-between-the-u-s-and-china-in-q4-2020.

[10] Forecast data is available on Foretell associated with two questions, both conditional on different election outcomes. See https://www.cset-foretell.com/questions/78-conditional-on-trump-winning-the-election-what-percentage-of-o-visas-will-go-to-chinese-nationals-in-fy-2021-october-1-2020-to-september-30-2021; https://www.cset-foretell.com/questions/79-conditional-on-biden-s-election-what-percentage-of-o-visas-will-go-to-chinese-nationals-in-fy-2021-october-1-2020-to-september-30-2021. Data on a related question covering an earlier time period is also available, https://www.cset-foretell.com/questions/47-what-percent-of-o-1-visas-issued-between-june-1-and-september-30-2020-will-go-to-people-from-china.

[11] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/80-what-percentage-of-u-s-citizens-will-have-an-unfavorable-view-of-china-as-reflected-in-the-2021-pew-global-attitudes-survey.

[12] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/32-how-many-sorties-against-chinese-aircraft-will-the-japanese-air-self-defense-force-conduct-by-april-30th-2020.

[13] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/61-how-much-will-the-u-s-department-of-defense-spend-on-ai-research-contracts-between-july-1-and-december-31-2020-inclusive.

[14] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/40-how-much-will-dod-spend-on-ai-grants-between-june-1-and-december-31-2020.

[15] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/74-what-will-the-combined-revenue-of-alphabet-amazon-apple-facebook-and-microsoft-be-in-the-first-two-quarters-january-1-through-june-30-of-2021. Data on a related question covering an earlier time period is also available, https://www.cset-foretell.com/questions/63-what-will-the-revenue-of-alphabet-amazon-apple-facebook-and-microsoft-be-in-the-third-quarter-of-2020-ending-september-30.

[16] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/69-how-much-new-funding-will-private-u-s-tech-companies-raise-between-july-1-and-december-31-2020-inclusive.

[17] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/82-how-many-postings-for-u-s-jobs-requiring-machine-learning-skills-will-be-published-between-january-1-and-march-31-2021-inclusive. Data on a related question covering an earlier time period is also available, https://www.cset-foretell.com/questions/19-how-many-new-u-s-jobs-requiring-machine-learning-skills-will-be-advertised-in-q3-2020.

[18] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/23-how-many-new-h-1b-applications-submitted-by-the-big-5-tech-companies-will-the-u-s-grant-in-fy-2020.

[19] Forecast data is available on Foretell, https://www.cset-foretell.com/questions/72-what-percentage-of-software-engineer-job-postings-between-january-1-and-march-31-2021-inclusive-will-allow-for-remote-work. Data on a related question covering an earlier time period is also available, https://www.cset-foretell.com/questions/65-what-percentage-of-software-engineer-job-postings-between-july-1-and-september-30-2020-inclusive-will-allow-for-remote-work.

[20] See https://github.com/georgetown-cset/public-foretell.

[21] See the examples in the introduction, notes 1-3.

[22] See, e.g., Naomi Klein, "Screen New Deal: Under Cover of Mass Death, Andrew Cuomo Calls in the Billionaires to Build a High-Tech Dystopia, Intercept, May 8, 2020, https://theintercept.com/2020/05/08/andrew-cuomo-eric-schmidt-coronavirus-tech-shock-doctrine/. To see how Klein's forecast can be operationalized on Foretell, see Maia Nikoladze, "Tech Companies Threaten Democracy," Foretell Blog, July 30, 2020, https://www.cset-foretell.com/blog/tech-threat.

[23] See, e.g., Cady, Etzioni, "China May Overtake US in AI Research," Medium, March 13, 2019, https://medium.com/ai2-blog/china-to-overtake-us-in-ai-research-8b6b1fe30595.

[24] See, e.g., Castro et al., "Who Is Winning the AI Race: China, the EU, or the United States?," Center for Data Innovation, August 19, 2020, https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states.

[25] CSET works with Cultivate Labs to recruit and maintain a pool of forecasters. To learn more about Foretell's forecasters, see our fall cohort of forecaster ambassadors at https://www.cset-foretell.com/our-ambassadors.