Automated Forecasts of Catastrophic Risks

Jason Abaluck* Ezra Karger* Nick Merrill*
Philip E. Tetlock* Eva Vivalt* Bridget Williams*

September 2026

Abstract

People hold radically different views about the risk of global catastrophe from AI. We present evidence that AI systems can help adjudicate these views: frontier models now predict binary real-world events about as accurately as the best humans ("superforecasters") and predict rare events in complex simulations increasingly well as AI capabilities improve (Spearman $\rho = +0.85$). We introduce the Automated AI Risk Outlook (AIRO), a regularly updated panel of frontier models' forecasts of catastrophic risk. At time of writing, the panel's median forecast that AI causes a catastrophe leading to the death of at least 10% of humans is 0.47% by 2030, 6.0% by 2050, and 12.2% by 2100. The AI-catastrophe median is about 45% of the all-cause median by 2030 and 71% by 2050. We also elicit cumulative incident forecasts on a standardized death-and-damage scale. At the highest threshold (equivalent to one billion deaths), misalignment is the highest risk, followed by cyber and biological risks. Estimates of risk are contingent on AI capabilities: conditioning on each model's own 90th-percentile estimate of frontier capability by March 2027 raises its AI-catastrophe forecast by 2030 by 2.21×. Models forecast that compute caps and pre-release reviews would meaningfully reduce risk, although these policies would have other costs that we do not model. Because our panel's accuracy rises with the same capability growth that likely drives potential AI risks, these forecasts are most informative in precisely the future worlds where AI risk is highest.

*Authors listed alphabetically. Jason Abaluck: Yale University and the National Bureau of Economic Research; Ezra Karger: Federal Reserve Bank of Chicago; Nick Merrill: Forecasting Research Institute and University of California, Berkeley; Philip E. Tetlock: University of Pennsylvania; Eva Vivalt: University of Toronto; Bridget Williams: Forecasting Research Institute and University of Oxford. This work was funded by a grant from Coefficient Giving. The views expressed in this paper do not necessarily represent the views of the Federal Reserve Bank of Chicago or the Federal Reserve System.

1 Introduction

People differ wildly in their assessment of catastrophic risks posed by AI. AI leaders like Dario Amodei and Elon Musk estimate a 10–25% chance of catastrophic outcomes including human extinction within the next few decades (Morrone, 2025; Tangalakis-Lippert, 2024), while academic experts express a wide range of views; Geoffrey Hinton (2018 Turing Award; 2024 Nobel Prize in Physics) gives similarly high estimates of extinction risk (Milmo, 2024) while Yann LeCun (2018 Turing Award) places this probability below that of extinction from an asteroid hitting the earth ($< 0.001\%$ over the next millennium) (LeCun, 2023). Surveys of AI experts and superforecasters in 2022–2023 placed the probability of AI-caused extinction by 2100 at 3–5% and 0.4%, respectively (Karger et al., 2023; Grace et al., 2025), although both groups underestimated subsequent AI capability progress (Kučinskas et al., 2025). If true risk is substantial, mitigating it could be the single most important public policy priority in human history, especially if this could be done without sacrificing the substantial possible benefits of AI (Jones, 2024). On the other hand, if the lowest-end estimates are right, costly slowdowns proposed by many lab leaders and pundits would achieve little while stymying potential scientific innovation and economic growth.

We believe this debate is missing a key input: forecasts from frontier large language models. The best models now approach superforecaster levels in the accuracy of their predictions; models are rapidly becoming more capable, and more capable models make better forecasts (Karger et al., 2025; ForecastBench, 2026). We show below that more capable models also make better forecasts of low-probability events in simulated environments. This evidence does not prove that more capable models can accurately forecast catastrophic risks, but it is suggestive. If models become more capable in ways that lead to increased risks from AI, model forecasting should likewise improve, providing increasingly accurate warning signs.

If the probability of catastrophe does spike and regulators or policymakers determine that coordination and decisive action are needed, accurate forecasts could act as a bellwether, alerting frontier labs and governments about which actions would most clearly mitigate risk. Forecasts could also act as a Schelling point, helping to solve coordination problems between competing AI companies or countries by establishing common knowledge of status quo risks and policy counterfactuals.

We introduce the Automated AI Risk Outlook (AIRO), a dashboard showing frontier model forecasts of various types of catastrophic risk. Our main definition of a catastrophe is an event that leads to the death of at least 10% of the global population within a short period of time (five years). Our dashboard includes top-line forecasts of overall catastrophic risk and catastrophic risk from AI specifically. It also includes severity-threshold forecasts for AI-related biosecurity, cybersecurity, and misalignment incidents. In the domain-specific forecasts, each threshold can be reached through deaths (or equivalent morbidity) or economic damages. We report separate forecasts from various frontier models, as well as an ensemble forecast that subsets to models with the highest score on the Epoch Capabilities Index (ECI), a measure of general AI capabilities that strongly correlates with real-world forecasting ability.

Figure 1 shows two views of the dashboard: forecasts of AI catastrophe over time and forecasts of frontier AI capability alongside human forecasts and a trend projection.

Our top-line findings are:

1. The ensemble of models places the probability of any catastrophic event (AI or non-AI) at 1.1% by 2030 (models span 0.7–1.6%), 8.5% by 2050 (3.2–11.0%), and 18.5% by 2100 (7.8–28.0%); the corresponding AI-catastrophe medians are 0.47%, 6.0%, and 12.2%, respectively.

2. At one billion deaths or equivalent morbidity, misalignment has the highest median among the three domain-specific categories at the 2030, 2050 and 2100 horizons (0.29% by 2030, 4.8% by 2050, 9.1% by 2100). Cyber incidents and AI-related human-caused epidemics are nearly tied at this threshold in 2030 and 2050; by 2100, their medians are 2.5% and 2.2%, respectively.

3. The ensemble median probability of an AI-caused catastrophe by 2030 falls from 0.47% unconditionally to 0.21% when each model conditions on its own 10th-percentile estimate of frontier capability in March 2027, and rises to 1.2% at its own 90th percentile. The medians of these low and high ECI estimates are 170 and 188, respectively. In other words, if AI capabilities decelerate relative to model expectations, models predict that risk would fall. If they accelerate relative to model expectations, risk would rise significantly.

4. International compute caps reduce the AI-catastrophe forecast by 38% by 2030 and 41% by 2050; international pre-release requirements reduce it by 44% and 47%, respectively. US-only versions have smaller predicted effects. We do not weigh these risk reductions against potential costs from foregone innovation and growth.

Our goal is both to report baseline numbers and to monitor how such forecasts evolve over time in response to new information. A virtue of model-based forecasting is that it accommodates arbitrarily granular questions without inducing survey fatigue, and so the above forecasts can also be elicited for a broad range of timelines and severity levels. We report these results below.

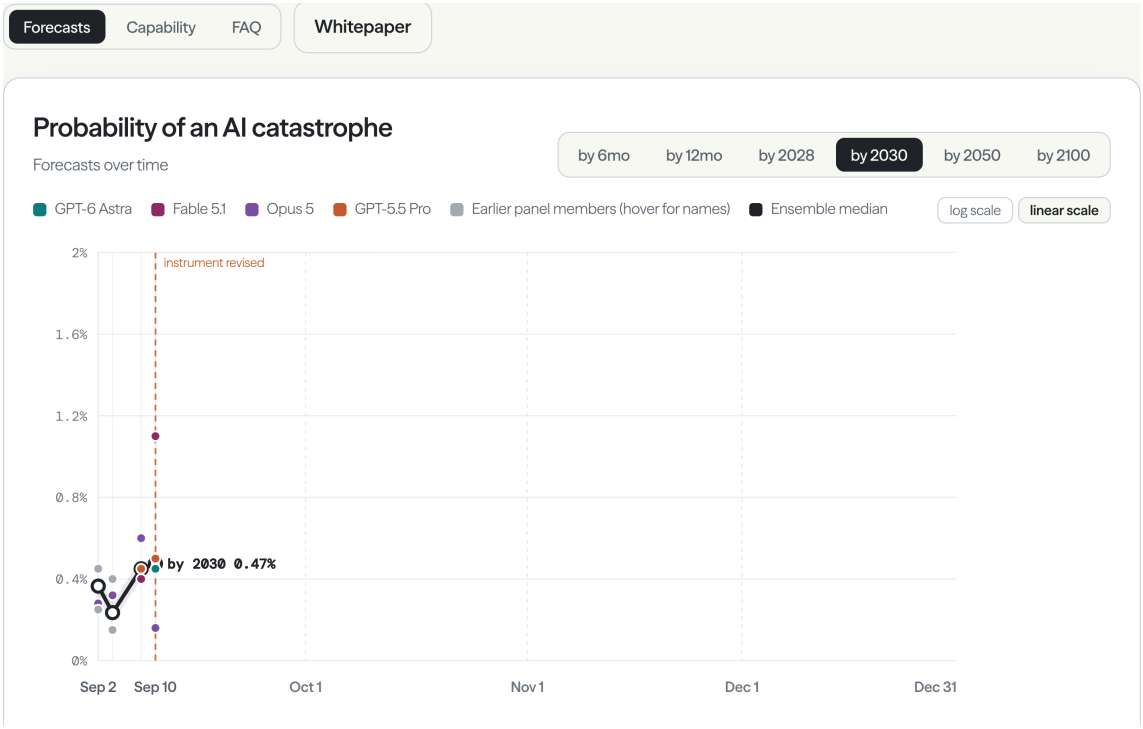
Our top-line forecasts are unconditional forecasts, meaning they incorporate any anticipated policy responses. We report preliminary evidence of models' conditional forecasting ability, and use policy-conditional forecasts to examine how much mitigation they "price in." For an AI-caused catastrophe by 2050, the ensemble median is 7.2% under the status-quo condition of no new substantial AI policy, versus 6.0% unconditionally. The unconditional median is thus 17% lower than the no-new-policy median, consistent with anticipated policy responses reducing risk.

These results raise two questions: should we believe them, and if so, how should policymakers respond?

The main reason to give these results some credence is that frontier models are now good at forecasting and, specifically, they are good at forecasting low-probability events. Our ForecastBench analysis finds less favorite-longshot bias and similar overall Brier scores for the evaluated models relative to the 2024 superforecaster panel (Figure 2); although, the question sets and evaluation dates differ.[1] We also apply techniques from Lee et al. (2026) to generate a large number of contamination-free, low-probability events in simulated video game worlds. Among these low-probability events, we find that model capability strongly predicts forecasting accuracy and that the most capable models' forecasts (including every model in our ensemble) approach

[1]The benchmark and its current leaderboard provide additional context for this comparison (Karger et al., 2025; ForecastBench, 2026).

(a) Forecast history and horizon selection



(b) Frontier capability and comparison forecasts

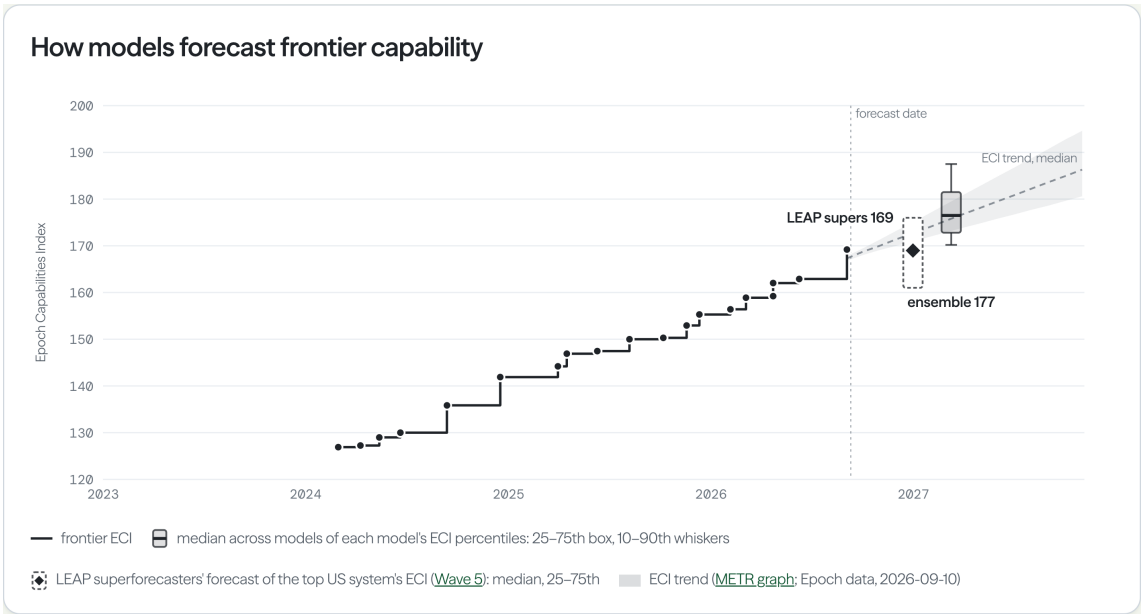


Figure 1: **The AIRO dashboard.** (a) Model-level and ensemble forecasts of AI catastrophe, with the 2030 horizon selected. (b) Frontier Epoch Capabilities Index (ECI), model forecasts, LEAP superforecaster estimates (Murphy et al., 2025), and an ECI trend projection. Screenshots captured in September 2026, using the September 2026 forecasts.

the true base rate of low-probability events.[2] We also show below that more capable models update more in response to relevant information they did not previously possess (market prices for off-line models), but update less in response to being prompted to read publicly available information, as we would expect if forecasts efficiently incorporated this information already. We expect continued improvements in model capabilities to create opportunities for better forecasting, but more testing is needed to validate model performance on real-world low-probability events, as well as specific precursors of catastrophic risk.

What should policymakers do with this information? Even small catastrophe probabilities imply substantial potential harm. The level of risk in our current ensemble forecast translates to at least 3.8 million statistical deaths by 2030. Even a small reduction in catastrophe probability could have substantial benefits.

To understand the possible benefits of approaches to mitigating AI risk, we consider *counterfactual* forecasts: predictions of risk outcomes conditioned on a specified policy being adopted. We validate this exercise in a simulated epidemic with known intervention effects, where frontier models recover most of the true effect (median 0.78 of the recoverable skill) and do so more accurately as AI capabilities rise. We then return to real-world policy counterfactuals and ask models to forecast AI risk conditional on 8 policy scenarios adapted from the forthcoming August 2026 wave of the Longitudinal Expert AI Panel (LEAP) Wave 12 ("Safety Policies"; Forecasting Research Institute, 2026). The model ensemble forecasts that federal pre-emption of state AI regulation increases the forecast probability of an AI-caused catastrophe 1.24× by 2050. An international compute cap, international pre-release authorization, and strict liability each lower forecasts of risk, and implementing the combined package reduces the forecast of AI-catastrophe by 66% by 2050 (0.34×). None of these policies is necessarily 'free': a full analysis would have to weigh forgone innovation against the risk reduction, which we do not attempt here. The conditional forecasts do provide a ranking of which policies are predicted to reduce risk the most. The panel expects international measures to be more effective than national ones, a bundle of measures to be more effective than single measures, and federal preemption to increase risk. We will continue to update this panel as new information arrives and as new policies are proposed.

The remainder of the paper proceeds as follows. Section 2 describes how we formulated our forecasting questions and elicited forecasts. Section 3 presents the evidence that our model forecasts are well-calibrated: on resolved real-world questions from ForecastBench, on low-probability events in simulated worlds, and on causal conditionals. Section 4 reports the panel's unconditional forecasts and the policy-conditional forecasts. Section 5 discusses what we have learned from calibrating the ensemble and how our results should be interpreted; Section 6 examines potential limitations to this approach, and Section 7 concludes.

2 Producing Forecasts

2.1 Question Formulation

To generate the questions for this exercise, we draw on earlier work by researchers at the Forecasting Research Institute (FRI). Using question formulations from prior studies lets us compare AI model forecasts with those of human forecasters, including subject matter experts and superforecasters. Specifically, we reuse question formulations on the risk of catastrophe and AI-related

[2]We are performing additional tests of these claims, which we plan to release in the near future. We plan to elicit model-based forecasts of both real-world low-probability events as well as intermediate events related in various ways to AI progress and risks (such as model capability progress).

catastrophe from the Existential Risk Persuasion Tournament (conducted in 2022; Karger et al., 2023) and we adapt question formulations from more recent work on biosecurity (Williams et al., 2026) and cybersecurity risks from AI (Ceppas de Castro et al., 2026). We updated these questions to suit the goals of this dashboard (e.g., removing a requirement that AI involvement occur in the year prior to the event) to better capture current thinking on AI risks. However, we think comparing these updated questions to human forecasts on very similar questions is still worthwhile.

We used the policy descriptions from a forthcoming survey of the FRI Longitudinal Expert AI Panel (Forecasting Research Institute, 2026). We developed these policy descriptions from a survey of AI policy proposals published in think tank reports and measures proposed by US politicians. We developed a long list of published policy proposals and, with input from external AI policy experts, refined it to six policies considered plausible and potentially impactful on AI risk.

We also added new questions where prior human forecasting left gaps in the discussion of mechanisms. These new questions include the probability of human disempowerment from AI and incidents arising from AI misalignment. We consulted with a superforecaster to ensure that all questions were precise enough to allow meaningful forecasting.

All four incident categories—any AI-related incident, AI-related human-caused epidemic, AI-related cyber incident, and misaligned AI incident—use the same eight severity thresholds. Each is defined by deaths (or equivalent morbidity) *or* economic damages, with $2.2 million per death-equivalent; for example, 100,000 deaths or $220 billion. Appendix A reproduces the full resolution criteria.

2.2 Eliciting Model Forecasts

We elicit forecasts from an ensemble of the top four highest-scoring models on the Epoch Capability Index (ECI), skipping within-family models (e.g., if Fable 5.1 is ranked second and Fable 5 is ranked fourth, we use only Fable 5.1). As of the time of this writing, these models are GPT-6 Astra, Fable 5.1, Opus 5, GPT-5.5 Pro. For each unconditional question and horizon, the ensemble forecast is the unweighted median of the four model probabilities. For conditional forecasts, we first divide each model's conditional probability by its own same-session unconditional probability, then take the median of the log ratios and exponentiate. With four models, this is the geometric midpoint of the two middle numbers. When we talk about ensemble multipliers of unconditional risk below, we are referring to the relative increase or decrease in risk implied by the ensemble forecast, calculated across top models.

We provide each model with the full survey instrument in a single prompt of 35 questions grouped by cause of catastrophe, each with its own resolution criteria and the horizons it is asked over (210 question–horizon cells), followed by the 14 conditions under which every cell is forecast (unconditional, 8 policy conditions, and 5 capability conditions). Each model answers the questions in one forecasting session, which can include multiple research and submission phases. In initial experiments, asking related questions together raised internal coherence without reducing accuracy. A system prompt asks for calibrated, evidence-grounded probabilities, and the harness withholds the submission tool until at least 10 web searches or page reads have been completed. Appendix A reproduces the system prompt, the framing of the user prompt, and representative question and condition blocks verbatim.

Following the prompt, each model conducts its own research via an agentic harness: a prompt asks it to investigate recent developments, expert reports, and published estimates for each cause

of catastrophe, read important sources, and use what it learns to formulate follow-up queries. The model chooses searches, receives results, and decides what to investigate next. We produce the elicitation prompts, forecasts, shared rationales, and model-gathered evidence in the raw records available through the dashboard's "Download Forecasts" feature. All models receive the forecasting date and questions and use the same web-search and page-reading tools, supplied via Tavily. Models do not have access to their previous forecasting sessions.

3 Validating Model Forecasts

Since our forecasts resolve at timescales ranging from a few months to decades, we cannot score them directly. We take three complementary approaches to assessing whether our models are calibrated on questions of this kind: real resolutions from ForecastBench, simulated worlds whose rare events resolve by the thousand, and tests of how updating and anchoring behaviors vary with model capability.

ForecastBench. ForecastBench (Karger et al., 2025) asks models and, in one 2024 round, superforecasters, to forecast on overlapping subsets of real-world forecasting questions. While the 2024 superforecaster questions largely resolved prior to this project's release, we can use our models' most recent ForecastBench resolution data to explore calibration and accuracy of model and human forecasts. Figure 2 plots each forecaster's own deciles of predictions against observed frequency. Vertical displacement from the diagonal is *miscalibration*; horizontal spacing is *resolution* (these components are terms of the Brier score). Despite high overall accuracy, superforecasters show some favorite-longshot bias, with low-probability events occurring less frequently than they forecast and high-probability events occurring more frequently. Recently released models appear to show less favorite-longshot bias and have overall Brier scores similar to human superforecasters.

Simulated worlds. Some of the catastrophic risks we forecast are low-probability events according to models. While ForecastBench gives us real-world question resolutions, the data to date contain too few low-probability events to reliably assess forecasting capability, especially for more recent models; by nature, low-probability events are scarce in the historical record. To address this, we use two simulators from FBSim (Lee et al., 2026), a simulated-world forecasting benchmark that can generate questions of arbitrary true probability. Our first simulator is a FreeCiv (Civilization-style) game world, with 25,919 binary questions (e.g., "Will Nuclear Fission be discovered by turn 70?") all falling in the 1–9% true probability range. Our second simulator is StarSim, an agent-based epidemic simulator, in which cumulative-infection threshold questions fall into the 1–5% base-rate range. For each simulator, models see only the question and a summary of the state of the world (a 'world report') at the snapshot turn or day; they never see the base rates, so the task is partly to estimate them. We score accuracy against an oracle that sees all simulations in the corpus except the one it forecasts, giving the oracle model a good sense of base rates. We believe this oracle would be very challenging for any forecaster to beat; events are rare enough that estimating their true probability requires forecasters to engage with and reason about game mechanics. We score 16 models spanning Epoch Capabilities Index 114–163. We find that accuracy is strongly associated with model capability, with the most capable models approaching the true base rate of questions (Figure 3).

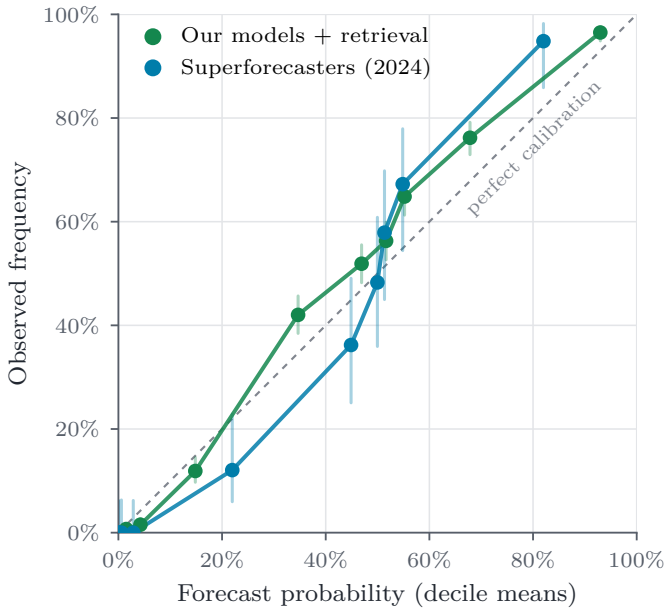


Figure 2: **Calibration on ForecastBench.** Observed frequency against forecast probability, in each forecaster's own deciles, with 95% Wilson intervals.

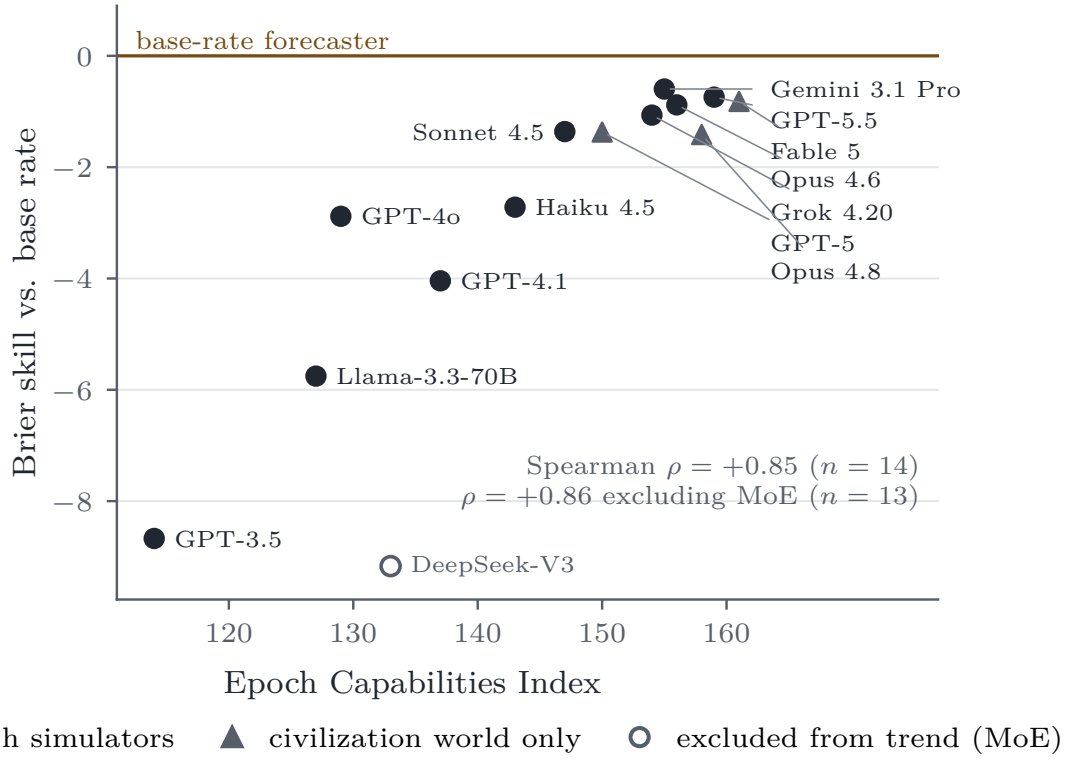


Figure 3: **Low-probability forecasting skill.** Brier skill score relative to the base-rate forecaster on low-probability questions from two simulated worlds with known generative processes and no contamination. (On low-probability events, estimating the true base rate is challenging.)

Causal conditionals. The policy forecasts in Section 4 ask a model to forecast events in response to a hypothetical intervention. To understand how well-calibrated our models are in causal forecasting, we use the same epidemic simulator we used for low-probability events: given a stated world and its observed trajectory, models forecast new infections *with* and *without* a vaccination campaign, in separate prompts, and we compare the true distribution from simulations run over different random seeds to the forecasts of models (Figure 4). Frontier models recover most of the recoverable skill under intervention (median 0.78 of the distance to the simulator's distribution, and the skill rises with capability (Spearman $\rho = +0.68$, $n = 23$, $p = 0.0003$), whereas the unconditional baseline does not ($\rho = +0.27$, $p = 0.21$) (i.e., the harder the causal question, the more capability matters).

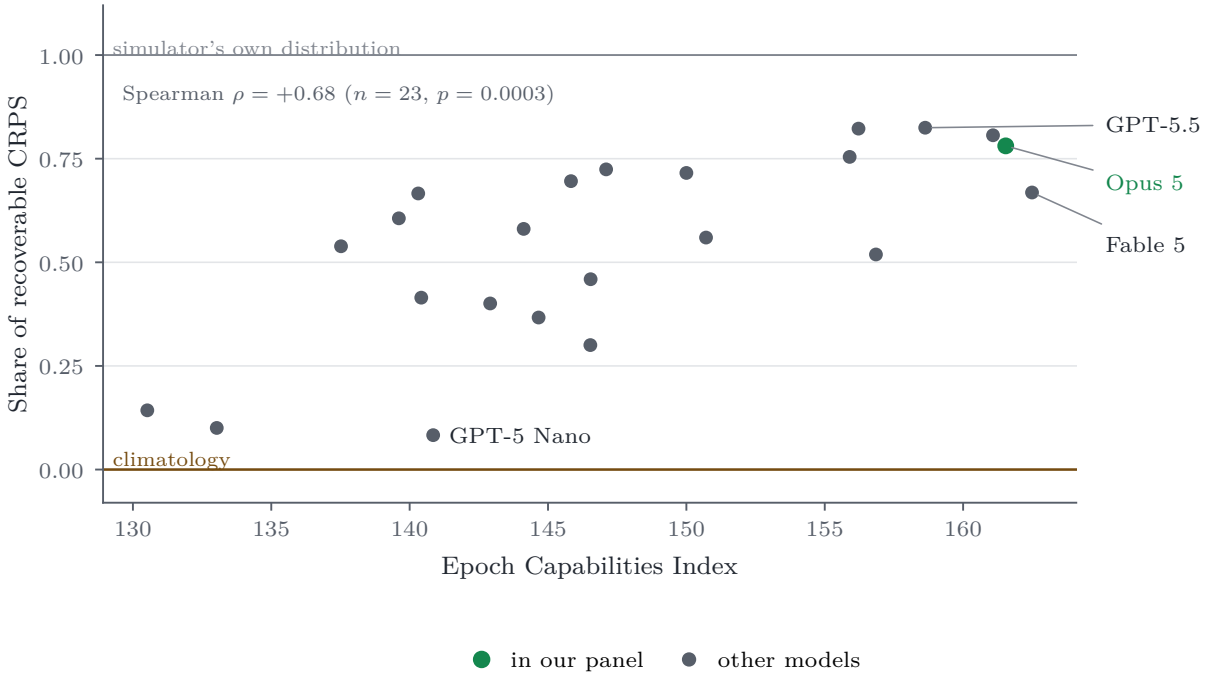


Figure 4: **Causal-conditional skill against capability.** Share of recoverable CRPS that each model's intervention forecasts recover, pooled over three vaccine-coverage rungs (25/50/90%) in a single-population SIR simulator with stated parameters: 1 matches the simulator's distribution, 0 matches base-rate forecasting.

Model anchoring and updating Our forecasting protocol is designed to make sure models review relevant information before making forecasts. This raises two questions: are models properly updating to incorporate new information, and are they overreacting to these readings, potentially anchoring on what they saw most recently?

For the first question, we find that better models update more in response to relevant information and, as a result, make more accurate forecasts. We use historical ForecastBench forecasts elicited with and without a market or crowd probability (Karger et al., 2025). For each model, we measure how much models update towards the market forecast by regressing the change in the model forecast on the initial gap between the model and market forecast, with no intercept. A value of zero indicates no directional movement toward the market probability; one indicates

closing the gap on average under this weighting. Figure 5, panel A, shows that more capable models move further toward the market probability across 64 model versions using all available forecast rounds, whose questions and evaluation dates differ (Spearman $\rho = 0.41$; slope per ten ECI points = 0.059, HC3 standard error = 0.020). Models scoring ten ECI points higher close an additional 5.9% of the initial model–market gap, on average. Restricting the comparison to 19 models answering the same 153 questions in April 2026 yields a similarly high correlation (Spearman $\rho = 0.62$; slope per ten ECI points = 0.158, HC3 standard error = 0.071). For 63 of 64 model versions, Brier scores were lower in forecasts made after exposure to the market price.

For the second question, we asked models to read fixed passages from six authors making AI-risk-related arguments, selected to present a range of perspectives: Yann LeCun, Gary Marcus, Toby Ord, Peter Wildeford, Geoffrey Hinton, and Eliezer Yudkowsky. We then asked models to assess their claims and incorporate them into their forecasts. Both conditions receive the same frozen evidence packet; the treatment adds all six authors' passages to test whether additional AI-related information shifts forecasts. We test whether models update at all in response to these additional readings, and whether they update differently in response to the most recent reading. Figure 5, panel B, reports the average absolute change across 35 unconditional catastrophic-risk questions for 2100, comparing three baseline calls with three passage calls (in randomized orders). More capable models update weakly *less*, although the point estimate is not statistically significant. Additionally, models do not differentially update based on whether the last reading they were exposed to was pro- or anti-AI.

These patterns are consistent with a world in which forecasts by more capable models are more efficient. They more readily incorporate new, forecast-relevant information, but make only small updates in response to already available information and arguments.

4 Results

4.1 Unconditional forecasts

The incident ladders show how forecast risk varies with severity and timing. At the 2030, 2050, and 2100 horizons shown in Figure 6, cyber incidents have the highest medians among the three domain-specific categories at low thresholds, while misalignment has the highest median at the top rung. Cyber and biological risks are nearly tied at the top rung in 2030 and 2050, with cyber higher in 2100. This ordering does not hold at every shorter horizon: at six months, the biological median exceeds both other domain medians at the top rung.

Models put a 10%-of-humanity catastrophe from any cause at 1.1% by 2030 [0.7–1.6], 8.5% by 2050 and 18.5% by 2100 (Figure 7). The corresponding AI-catastrophe medians are 0.47%, 6.0% and 12.2% at the three horizons. The AI-catastrophe median is about 45% of the all-cause median by 2030 and 71% by 2050. The ensemble median for human disempowerment exceeds that for AI mortality catastrophe at every reported horizon (15.5% by 2050, 28.0% by 2100).

For both anchored questions, the model ensemble median exceeds the LEAP superforecaster median at all three shared fixed-year horizons (2030, 2050, and 2100). For the general question, the ratios range from 1.75× to 2.83×; for AI catastrophe, they range from 4.75× to 6.82×. These ratios are calculated before rounding the displayed probabilities; they compare forecasts made at different dates and do not measure relative accuracy. For the unconditional forecasts, across 1,996 implied orderings (a probability cannot fall as the horizon lengthens, the severity threshold loosens, or the cause set widens), 100.0% hold.

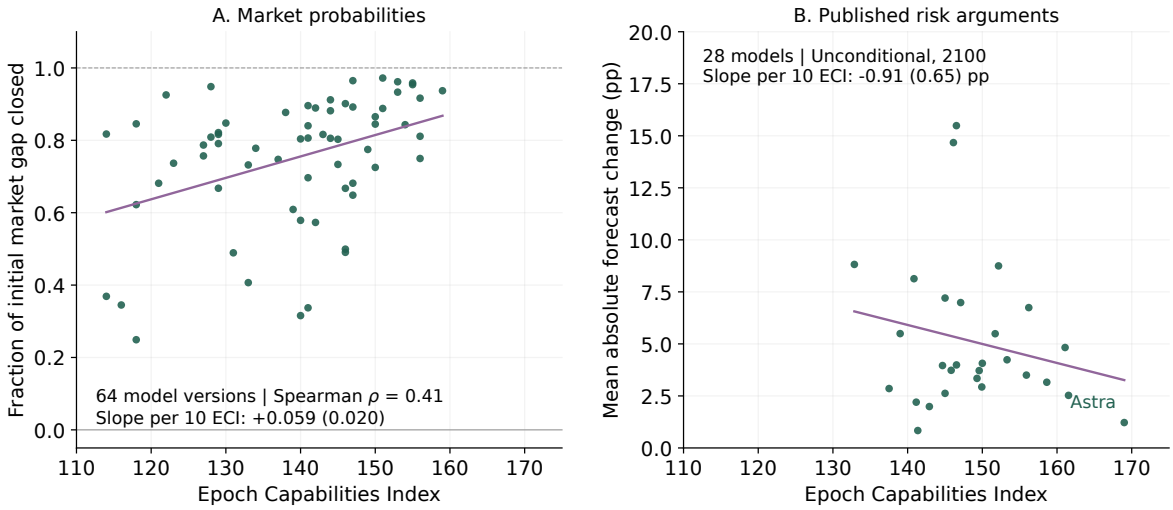


Figure 5: **Model capability and forecast updating.** Panel A: prediction-market probabilities. Fraction of the initial model–market gap closed after exposure to the market probability, across 64 model variants. Panel B: published catastrophic-risk arguments. Mean absolute change across 35 unconditional forecasts for 2100 after reading six authors, in percentage points, across 28 models. Both panels use the same ECI scale; their y-axes measure different quantities. Lines are unweighted linear fits with intercepts. Slopes are per ten ECI points, with standard errors in parentheses. The sets of models do not overlap because the Panel B analysis was run more recently, and many models have been deprecated.

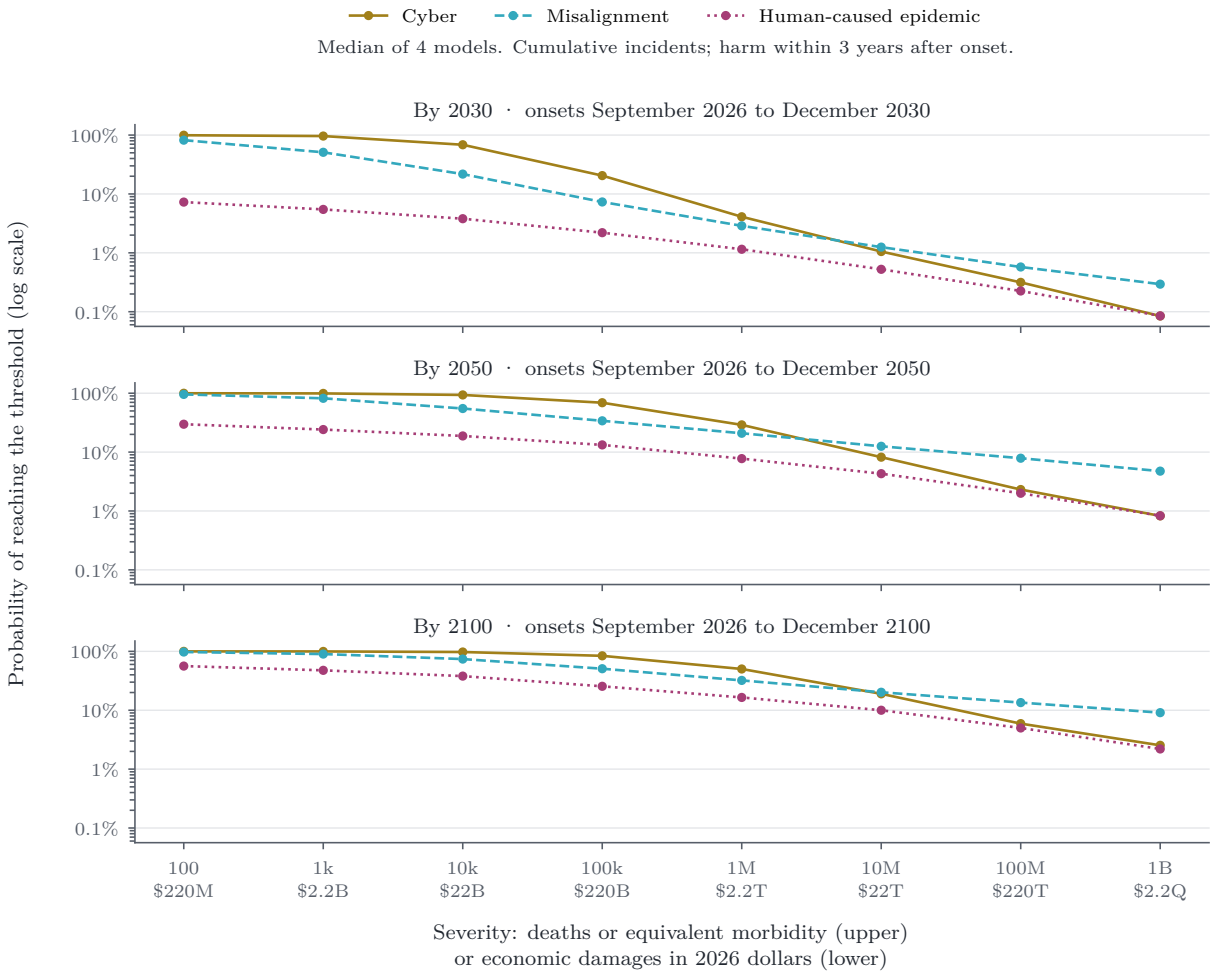


Figure 6: **Incident ladders.** Unconditional probabilities of cumulative qualifying AI-related cyber, misalignment, or human-caused epidemic incidents reaching each death-or-damage threshold, using the unweighted median of four models. Upper tick labels give deaths or equivalent morbidity; lower labels give economic damages in 2026 dollars. Either threshold qualifies. Incident onset must fall between elicitation and the stated deadline. Categories overlap, and their probabilities should not be summed.

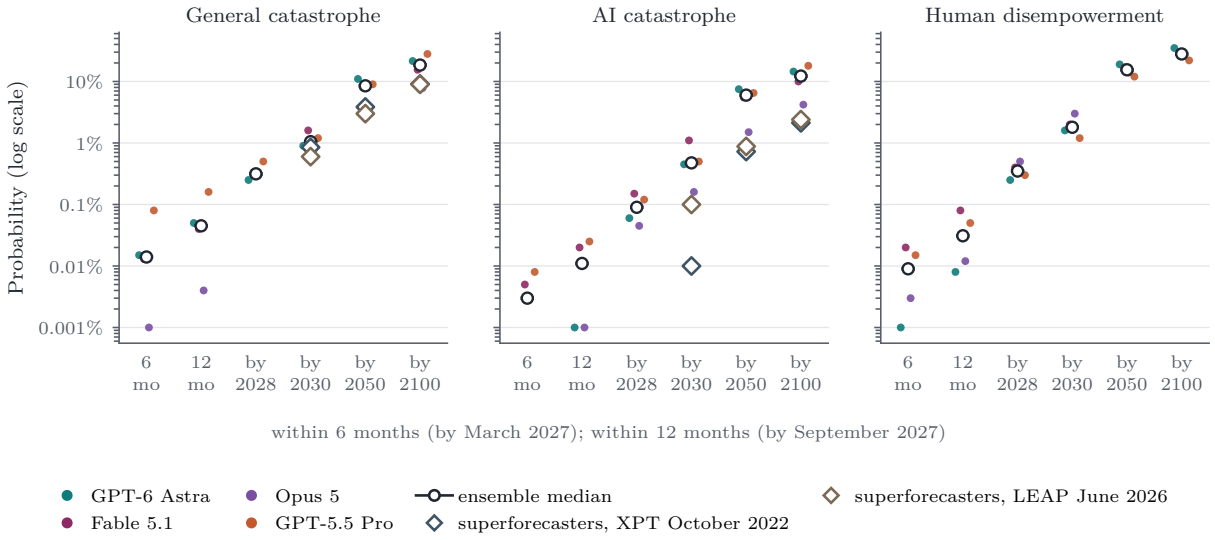


Figure 7: **Unconditional forecasts.** Probability that each event occurs within six or twelve months of the forecast, or by 2028, 2030, 2050, or 2100: a catastrophe causing the death of 10% of humanity from any cause, the same catastrophe caused by AI, and human disempowerment. Filled points are individual models; the open circle is the ensemble median. Diamonds are superforecaster medians from XPT (October 2022) and LEAP (June 2026); the disempowerment question has no human anchor. Log scale.

4.2 Conditional on policy

As described above (Section 2), we elicit forecasts on all risks conditional on a set of policies being implemented. For the six individual policy scenarios and the combined package, all 4 models agree on the direction of the AI-catastrophe effect at 2030 and 2050 (Figure 8).

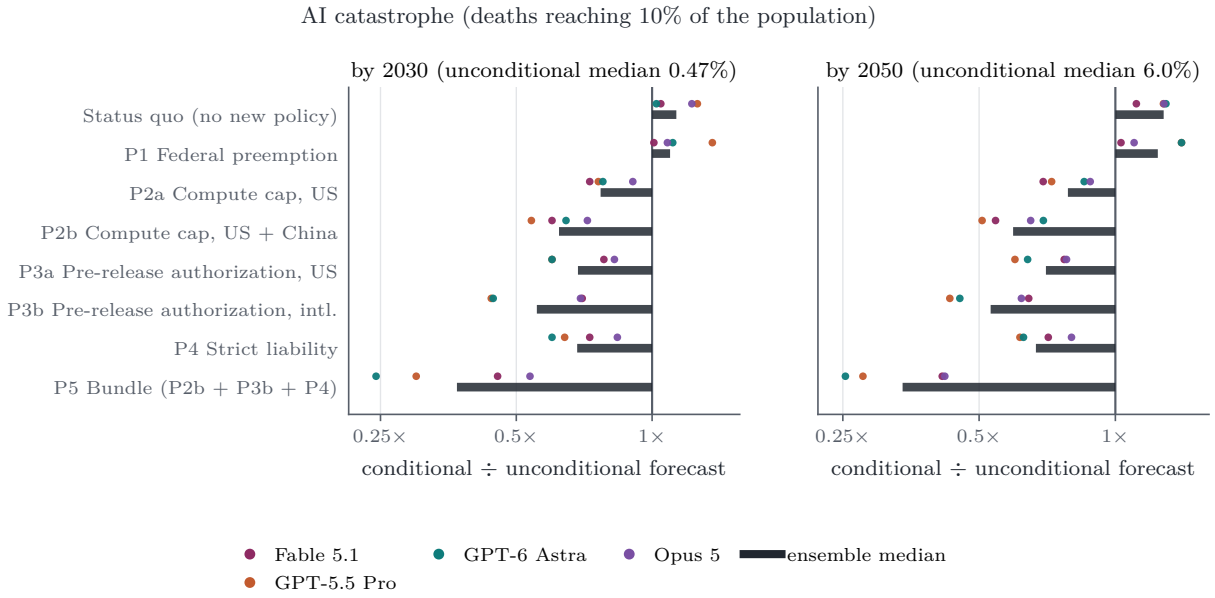


Figure 8: **Effect of policy on the AI-catastrophe forecast.** Bars summarize conditional-to-unconditional probability ratios using the median on the log scale; dots are the 4 models. Conditions are the eight LEAP Wave 12 policy scenarios.

The bundle (P5: an international compute cap, international pre-release authorization, and strict liability) reduces the AI-catastrophe forecast by 63% by 2030 and 66% by 2050, using the log-scale median of per-model ratios. The corresponding multipliers are 0.37× by 2030 [model range 0.24–0.54] and 0.34× by 2050. At 2050, international compute caps and pre-release authorization alone reduce forecasts by 41% and 47%, respectively; their US-only versions reduce forecasts by 21% and 30%. Strict liability alone reduces the forecast by 33%.

Two conditions raise risk forecasts: the status quo (1.28×) and federal preemption of state AI law (1.24×); both policies increase risk above the unconditional ensemble forecast at 2050. These two conditions imply that the ensemble's unconditional forecast "prices in" some degree of policy action.

4.3 Conditional on capability

We asked the ensemble to provide their estimates of the ECI achieved by the highest-ranking model in six months, eliciting the 10th, 25th, 50th, 75th, and 90th percentile estimates for each model in the ensemble. We then ask the models to forecast all risk questions conditional on achieving those ECI scores. As of writing, the ensemble puts the frontier at 176 in March 2027 (the medians of models' own 25th- and 75th-percentile estimates are 173 and 182). These summaries do not define common ECI conditions shared by all models: each model conditions

on its own percentile estimates. For comparison, the trend fitted to the September 2026 ECI snapshot projects 176 at the target date.

All models agree that higher model capabilities lead to elevated forecasts of risk. A world at the model's own 25th percentile lowers the AI-catastrophe forecast to 0.71× the unconditional by 2030 [0.60–0.82] and one at the 75th percentile raises it to 1.42× [1.24–1.70], with the same direction of effects by 2050 (0.76× and 1.29×). Across the 9-point interquartile range the forecast moves by 2.00×, about 8.3% per ECI point. The own-median condition produces an ensemble ratio of 0.99× at 2030, although individual models differ. Conditioning on a variable's median need not reproduce an unconditional probability.

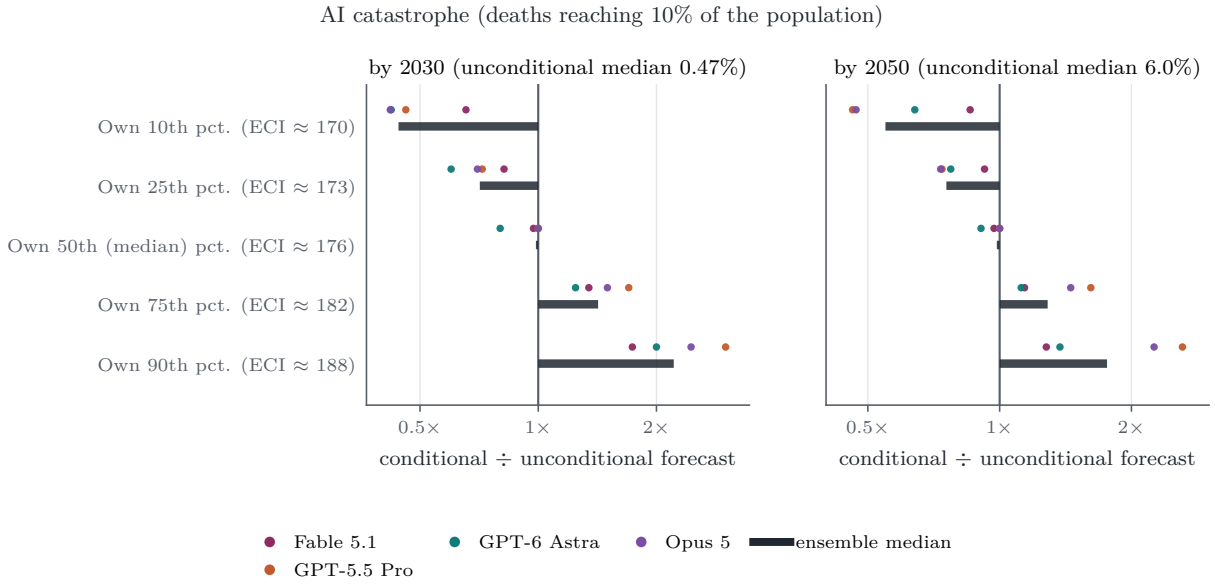


Figure 9: **Effect of frontier capability on the AI-catastrophe forecast.** Each model first forecasts the frontier ECI (Epoch Capabilities Index) in March 2027, 6 months after the run, giving its own 10th, 25th, 50th, 75th and 90th percentiles; each condition then fixes the index at one of those values (ensemble levels in parentheses) and asks the model to forecast in that world. Bars summarize per-model ratios using the median on the log scale; dots are the 4 models.

Additional conditional forecasts are presented in Figure 10. The high-growth and low-labor-force-participation scenarios generally have higher ensemble AI-catastrophe forecasts, consistent with models associating these outcomes with greater AI capability. This relationship is not uniformly monotonic: by 2030, the catastrophe median under −0.5% annual GDP growth is higher than under 1% growth. These conditional associations do not establish that faster growth itself causes catastrophic risk.

5 Discussion

Our results support the idea that AI forecasts can contribute to catastrophic-risk assessment. But, model accuracy on these questions is unknown, and the dashboard is intended to improve over time. Our forecasts depend on the trajectory of model capabilities, highlighting the potential benefits of better real-time measures of frontier model capabilities for monitoring risk.

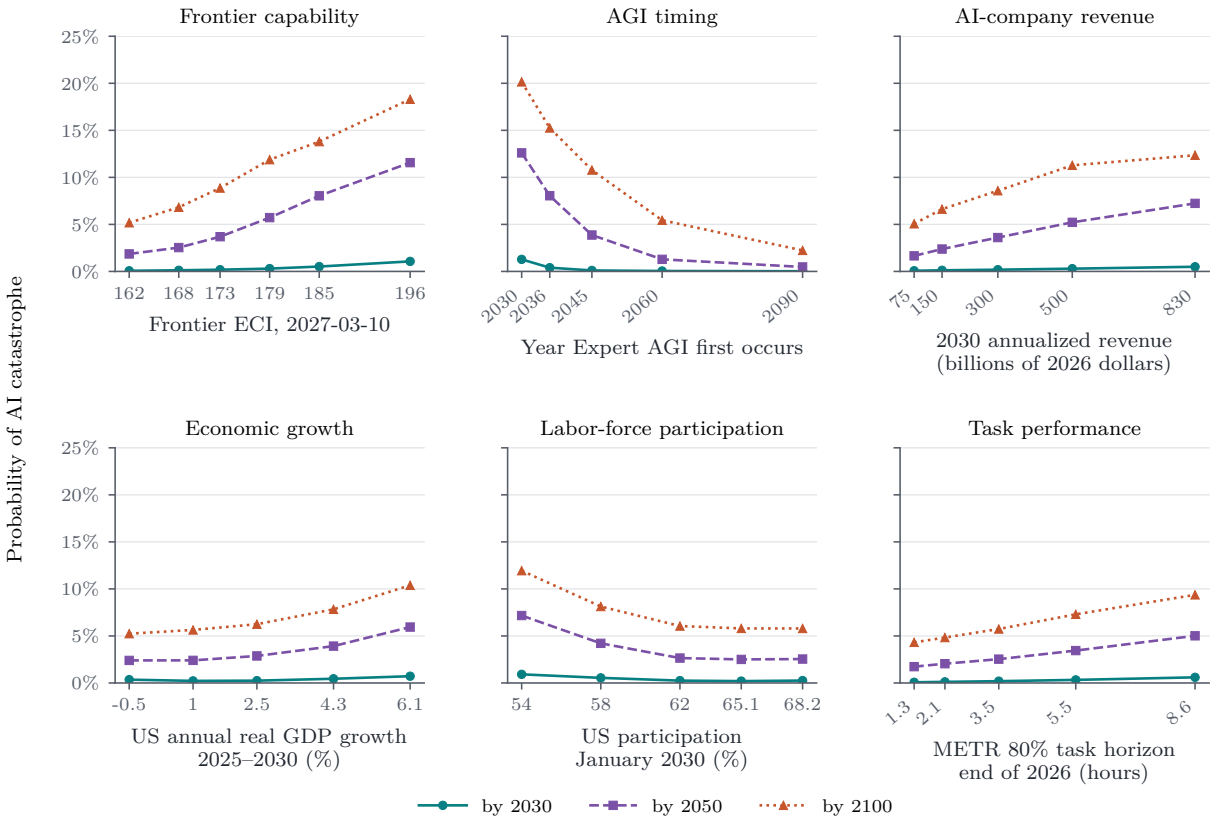


Figure 10: **Conditional risks of AI-related catastrophes.** Probabilities of AI-related catastrophes causing the death of at least 10% of the global population by 2030, 2050, and 2100, conditional on the outcomes shown in each panel, using the median of four models.

How do these forecasts compare to forecasts from humans? We have comparison data for some of the dashboard's questions, and we can also compare to prior work using related but different questions. The ensemble forecast of AI catastrophic risk produces consistently larger estimates than human AI experts who are asked the same question. In May 2026–June 2026, the median AI expert in the LEAP forecasted a 0.3% probability of AI causing the death of more than 10% of humans by the end of 2030 and 2% by the end of 2050, while our ensemble produced 0.47% by 2030, 6.0% by 2050. The median superforecaster put these figures lower, at 0.1% and 0.88% respectively.

Substantively, the models assign appreciable probabilities to some incident thresholds even over short horizons. The category ordering depends on severity and timing: cyber incidents lead at lower death-or-damage thresholds, while misalignment leads at the 1 billion+ deaths or equivalent after 2030. At six months, biological risk is most likely to lead to catastrophe.[3] Human-disempowerment medians exceed the AI mortality-catastrophe medians at every reported horizon, including six months.

The models forecast that several policies could mitigate risk, but policy responses must weigh these risks against the considerable potential benefits of further AI development.

Practically, we believe our dashboard can contribute to decision-making through a few channels:

Quantifying Risk Assessment Currently, AI companies assess risk using (often bespoke) benchmarks around critical areas like nuclear or biological risk, creating "redlines" that, if crossed, would justify withholding a model or delaying its release. Our forecasts can systematize these analyses. Conditional forecasts relate risk to benchmarks achieved by publicly released models. Thus, companies and regulators can determine the increase in risk and associated cost of missing a specific benchmark. AI companies can simulate how the ensemble updates forecasts conditional on a given model's performance. This approach makes it easier for companies to quantitatively determine risks incurred from new models and provides policymakers with an apples-to-apples comparison of risks over time and across companies and models.

Estimating the value of policies Our counterfactual forecasts provide policymakers with a common-unit benchmark for comparing policies in terms of effectiveness. Our ensemble can create an ordinal ranking of policies by efficacy. While our current, preliminary work looks only at the benefit of policy from catastrophe reduction, future work will also consider the downside to economic growth, life expectancy, or other measures. A future version of this general method could allow policymakers a more flexible "sandbox" for experimenting with and estimating the relative value of policies under consideration.

Determining what additional information might be most helpful Conditional forecasts can be used to identify areas of uncertainty and which outcomes have the highest value of information, to help prioritize monitoring and information-gathering.

A central aim of the project is to improve forecasts over time through repeated measurement, transparent definitions, and evaluation on questions that resolve sooner. Our forecasts will track

[3]The first severity rung at which misalignment exceeds cyber within six months is 100 million for GPT-6 Astra and GPT-5.5 Pro, and 10 million for Fable 5.1. Opus 5 has no such crossover among the eight rungs. Each rung refers to deaths or equivalent morbidity, or the corresponding economic damages.

changes over time, providing evidence for how catastrophic risk responds to events and mitigation efforts, including both capability increases and breakthroughs in model alignment techniques.

6 Limitations

This work has some important limitations. First, it is unclear how well our measures of forecasting accuracy — from ForecastBench and simulated worlds — correlate with forecasting accuracy on the questions we present here. ForecastBench measures performance on two types of binary questions: questions that resolve using time-series databases (such as inflation or pandemic deaths), and questions based on prediction markets and public forecasting platforms. For both question types, events are generally not rare, and forecasts usually have substantial time-series data to base them on, unlike forecasts of catastrophic events. These limitations are addressed by using simulated worlds to assess forecasting accuracy on low-probability events. However, the relationship between specifically low-probability simulated-world forecasting and low-probability real-world forecasting is not yet well-established. These proxies provide reasons to investigate model forecasts, but do not establish that frontier models are as accurate as the best human forecasters on the specific question of catastrophic risks from AI. We are developing additional methods to assess AI model accuracy across different types of forecasting and will update the dashboard as we develop more evidence.

Another limitation is that the dashboard's forecasts rely only on public information. Much information relevant to forecasting AI risks is private, including information from AI companies and classified national security information. In particular, frontier AI companies likely possess private information about capabilities and risks, so model forecasts would likely need access to proprietary information from companies to warn against potential risks arising from internal development and deployment (Abaluck, 2026).

Finally, the existence of this dashboard could itself make the forecasts less reliable. Publishing these forecasts means they may appear in the training corpus for future AI models, potentially reinforcing the forecasts. Furthermore, as models become more sophisticated, AI companies or models may collude or otherwise conceal data from automated forecasts, especially if policy responses are contingent on these forecasts. We can mitigate these risks, for example, by using suites of forecasting models that report reasoning quite transparently. We plan to explore these and other options for mitigating the risk of insincere forecasts and collusion in future work.

7 Conclusion

We report results from a new dashboard that tracks AI forecasts of catastrophic risk over time, where a catastrophe is an event whose attributable deaths reach at least 10% of the population alive at the start of the forecast window. Models place the risk of such a catastrophe by 2030 at 1.1% (models span 0.7–1.6%), rising to 8.5% by 2050 (3.2–11.0%). The corresponding AI-catastrophe medians are 0.47% by 2030 and 6.0% by 2050. Separately, at the highest death-or-damage incident rung, misaligned AI incidents have the highest domain-specific median at the 2030, 2050 and 2100 horizons. Cyber and biological risks are nearly tied at this threshold in 2030 and 2050, with cyber higher in 2100. These overlapping incident categories do not decompose the deaths-only catastrophe forecast. We also present new evidence that forecasts

from higher-capability models improve specifically for low-probability events, and that the best models exhibit less favorite-longshot bias than superforecasters.

Our dashboard provides information on catastrophic risks from AI to inform policymakers and regulators. The dashboard indicates how this risk is changing in response to new developments in AI progress and policy. It complements, not replaces, other approaches policymakers use to explore, understand, and mitigate AI risk.

Any forecast – including forecasts of rare events – can only be empirically validated by defining classes of related events and testing performance on those events. The difference between predicting catastrophic risk in 2100 and who will win the world series in 2026 is that we so far observe few indicators that we have a firm basis for thinking correlate with the realization of catastrophes in 2090. But it is still possible to construct and validate more such indicators given underlying assumptions about how events relate, and more feasible for shorter term predictions.

In future work, we plan to continue validating and extending our dashboard by creating ensemble forecasts and a larger set of intermediate questions related in various ways to endpoint catastrophic risk. These intermediate questions will both serve as validation tools and provide inputs that future models can use to inform their risk assessments. Finally, we will continue to augment the dashboard to include counterfactual forecasts to better understand the costs and benefits of proposed mitigation policies on risk.

References

Abaluck, J. (2026). *Conditional regulation of frontier AI with automated insider forecasts*. Working paper, June 2026. https://jabaluck.github.io/conditional-regulation/paper.pdf

Ceppas de Castro, R., Williams, B., Halstead, J., van der Merwe, M., & Connor, J. (2026). *Forecasting AI cyber risks and capabilities: Results of a 2025 pilot study*. Working paper no. 7. Forecasting Research Institute, July 23. https://forecastingresearch.org/research/ai-cyber-risks-capabilities

Dafoe, A. (2018). *AI governance: A research agenda*. Centre for the Governance of AI, Future of Humanity Institute, University of Oxford. Version 1.0, August 27. https://www.governance.ai/research-paper/agenda

ForecastBench. (2026). *Leaderboards*. Forecasting Research Institute. Accessed September 11, 2026. https://www.forecastbench.org/leaderboards/

Forecasting Research Institute. (2026). *Longitudinal Expert AI Panel, wave 12: Safety policies*. Internal survey and policy-description drafts dated August 2026. These drafts supply the scenarios used here; the fielded instrument may differ.

Grace, K., Sandkühler, J. F., Stewart, H., Weinstein-Raun, B., Thomas, S., Stein-Perlman, Z., Salvatier, J., Brauner, J., & Korzekwa, R. C. (2025). Thousands of AI authors on the future of AI. *Journal of Artificial Intelligence Research*, *84*, Article 9, 1–48. https://doi.org/10.1613/jair.1.19087

Halawi, D., Zhang, F., Chen, Y.-H., & Steinhardt, J. (2024). Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, *37*. https://doi.org/10.52202/079017-1598 arXiv:2402.18563

Jones, C. I. (2024). The AI dilemma: Growth versus existential risk. *American Economic Review: Insights*, *6*(4), 575–590. https://doi.org/10.1257/aeri.20230570

Karger, E., Monrad, J. T., Mellers, B., & Tetlock, P. E. (2021). *Reciprocal scoring: A method for forecasting unanswerable questions*. Working paper, October 31. SSRN 3954498. https://doi.org/10.2139/ssrn.3954498

Karger, E., Rosenberg, J., Jacobs, Z., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Williams, B., McCaslin, T., Thomas, S., & Tetlock, P. E. (2023). *Forecasting existential risks: Evidence from a long-run forecasting tournament*. Working paper no. 1. Forecasting Research Institute. Revised August 8, 2023. https://forecastingresearch.org/research/existential-risk-persuasion-tournament

Karger, E., Bastani, H., Chen, Y.-H., Jacobs, Z., Halawi, D., Zhang, F., & Tetlock, P. E. (2025). ForecastBench: A dynamic benchmark of AI forecasting capabilities. *International Conference on Learning Representations (ICLR)*. https://iclr.cc/virtual/2025/poster/28507 arXiv:2409.19839

Karnofsky, H. (2024, September 13). *If-then commitments for AI risk reduction*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction

Koh, A., & Sanguanmoo, S. (2026). *Technology speed limits*. arXiv:2606.01424, May 31.

Kučinskas, S., Rosenberg, J., Ceppas de Castro, R., Jacobs, Z., Canedy, J., Tetlock, P. E., & Karger, E. (2025). *Assessing near-term accuracy in the Existential Risk Persuasion Tournament*. Working paper. Forecasting Research Institute, September 2. https://forecastingresearch.org/research/near-term-xpt-accuracy

LeCun, Y. (2023). *Meta's chief AI scientist Yann LeCun talks about the future of artificial intelligence*. Video interview, CBS Mornings, December 16, 24:31–25:10. https://www.youtube.com/watch?v=Ah6nR8YAYF4

Lee, J., Merrill, N., & Karger, E. (2026). *ForecastBench-Sim: A simulated-world forecasting benchmark*. arXiv:2606.18686, June 17.

METR. (2025). *Common elements of frontier AI safety policies*. December 2025 version, December 16. https://metr.org/common-elements

Milmo, D. (2024). 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years. *The Guardian*, December 27. https://www.theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wiping-out-humanity-over-next-30-years

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). Position: Levels of AGI for operationalizing progress on the path to AGI. *Proceedings of the 41st International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, 235, 36308–36321. https://proceedings.mlr.press/v235/morris24b.html

Morrone, M. (2025, September 17). Amodei on AI: "There's a 25% chance that things go really, really badly." *Axios*. https://www.axios.com/2025/09/17/anthropic-dario-amodei-p-doom-25-percent

Murphy, C., Rosenberg, J., Canedy, J., Jacobs, Z., Flechner, N., Britt, R., Pan, A., Rogers-Smith, C., Mayland, D., Buffington, C., Kučinskas, S., Coston, A., Kerner, H., Pierson, E., Rabbany, R., Salganik, M., Seamans, R., Su, Y., Tramèr, F., Hashimoto, T., Narayanan, A., Tetlock, P. E., & Karger, E. (2025). *The Longitudinal Expert AI Panel: Understanding expert views on AI capabilities, adoption, and impact*. Working paper no. 5. Forecasting Research Institute, November 10. Includes results from waves 1–3. https://forecastingresearch.org/research/longitudinal-expert-ai-panel-leap-working-paper

OpenAI. (2025). *Preparedness framework* (Version 2, April 15). https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

Tangalakis-Lippert, K. (2024, March 31). Elon Musk says there could be a 20% chance AI destroys humanity—but we should do it anyway. *Business Insider*. Republished by Yahoo. https://tech.yahoo.com/ai/articles/elon-musk-says-could-20-235807723.html

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown. https://www.penguinrandomhouse.com/books/227815/superforecasting-by-philip-e-tetlock-and-dan-gardner/9780804136709/

UK Government, Department for Science, Innovation and Technology. (2024). *Frontier AI safety commitments, AI Seoul Summit 2024*. May 21; updated February 7, 2025. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

van der Merwe, M. (2026). *Could AI enable catastrophic cyberattacks on the US power grid?* Centre for the Governance of AI, July 30. https://www.governance.ai/research-paper/could-ai-enable-catastrophic-cyberattacks-on-the-us-power-grid

Williams, B., Righetti, L., Rosenberg, J., Ceppas de Castro, R., Kuusela, O., Britt, R., Soice, E., Morales, A., Sanders, J., Donoughe, S., Black, J., Karger, E., & Tetlock, P. E. (2026). *Forecasting LLM-enabled biorisk and the efficacy of safeguards*. Working paper. Forecasting Research Institute. Originally published July 1, 2025; revised April 30, 2026. https://forecastingresearch.org/research/llm-enabled-biorisk

Wilson, B. (2025a). *Q1 AI benchmark results: Pro forecasters crush bots*. Metaculus, June 21; updated October 13. https://www.metaculus.com/notebooks/38673/q1-ai-benchmarking-results/

Wilson, B. (2025b). *Q2 AI benchmark results: Pros maintain clear lead*. Metaculus, October 26; updated December 15. https://www.metaculus.com/notebooks/40456/q2-ai-benchmark-results/

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*(2), 107–126. https://doi.org/10.1257/0895330041371321

A The elicitation prompt

Protocol `unified-joint-combined-v5`, as sent on September 2026. The question set contains 35 questions; the four shown are representative, and every incident-ladder question shares the criteria block that follows them.

System prompt

```
You are a careful probabilistic forecaster producing calibrated
probabilities for catastrophic-risk questions.  Ground your forecast in
evidence, and gather it in steps:  web_search for current events, expert
reports, base rates and published estimates; read_page to read the results
that matter in full; then search again on what you learned.  Let each round
shape the next query.  Before you estimate, check what has happened recently
on each cause (web_search with recent_days, counted back from today's date
in your prompt).  The submission tools (submit_cells, submit_forecast) are
offered only after at least 10 tool calls have returned (searches or page
reads); stop once more searching would not move a number.  Reason from base
rates first, then adjust.  These are low-probability questions --- beware
both dismissing tail risk as zero and inflating it for vividness.  When
done, call submit_forecast.
```

User prompt

```
Today is 2026-09-10.

Below are 35 forecasting questions, grouped by cause of catastrophe.
Forecast all of them together, in one pass -- and forecast every question x
horizon cell under each of the 14 conditions listed in the CONDITIONS
section that follows the questions:  once unconditionally, and once under
each condition in each of its 2 groups (policy conditions and capability
conditions).  Every condition is taken on its own; no condition from one
group is combined with a condition from another.

--- catastrophe:general (severity:  deaths reaching 10% of the population)
---
What is the probability that any global catastrophe (deaths reaching at
least 10% of the population alive at the start of the measurement window)
will occur within 6 months / within 12 months / by the end of 2028 / by the
end of 2030 / by the end of 2050 / by the end of 2100?
Horizons for this question:  6mo, 12mo, 2028, 2030, 2050, 2100

Resolution criteria:
* Catastrophe threshold and the measurement window:  A catastrophe at the
10% level means that deaths attributable to the catastrophe (excess
mortality, i.e., deaths beyond the baseline that would have occurred anyway)
reach at least 10% of the population alive at the start of the measurement
window.
* The measurement window is any contiguous period of at most five years
during which the attributable deaths accumulate to the threshold.  The
window must begin on or after December 31, 2025 and end on or before
December 31 of the resolution year (2030, 2050, or 2100).  A faster event;
the five years is a ceiling on the span over which deaths are counted, not a
```

required duration.
* For a threshold to resolve positively, the deaths must arise from a single catastrophe or a single causally-linked chain of events, not from summing unrelated causes.
* Data and disagreement: For fatality and population data, we will rely on broadly recognized sources (e.g., Our World in Data, the Economist, the United Nations). Where these disagree in ways that would change resolution, we will resolve by the majority opinion of a panel of experts in demographics and fields relevant to the specific catastrophe (e.g., epidemiology for a pandemic). If reasonable people disagree about whether the event occurred, the question resolves via the same expert panel.
* [Scope of "the resolution year", added 2026-08-28 when these questions moved onto the incident ladder's horizon grid: read it as the resolution date of whichever horizon you are answering, which for a rolling horizon is that many months after today's date. The rule is otherwise unchanged -- the measurement window still begins on or after December 31, 2025, must end on or before that resolution date, and still spans at most five years.]

--- catastrophe:ai (severity: deaths reaching 10% of the population) ---
What is the probability that a global AI-related catastrophe (AI being the proximate cause, per the but-for standard, of attributable deaths reaching at least 10% of the population alive at the start of the measurement window) will occur within 6 months / within 12 months / by the end of 2028 / by the end of 2030 / by the end of 2050 / by the end of 2100?
Horizons for this question: 6mo, 12mo, 2028, 2030, 2050, 2100

Resolution criteria:
* Catastrophe threshold and the measurement window: A catastrophe at the 10% level means that deaths attributable to the catastrophe (excess mortality, i.e., deaths beyond the baseline that would have occurred anyway) reach at least 10% of the population alive at the start of the measurement window.
* The measurement window is any contiguous period of at most five years during which the attributable deaths accumulate to the threshold. The window must begin on or after December 31, 2025 and end on or before December 31 of the resolution year (2030, 2050, or 2100). A faster event; the five years is a ceiling on the span over which deaths are counted, not a required duration.
* For a threshold to resolve positively, the deaths must arise from a single catastrophe or a single causally-linked chain of events, not from summing unrelated causes.
* Data and disagreement: For fatality and population data, we will rely on broadly recognized sources (e.g., Our World in Data, the Economist, the United Nations). Where these disagree in ways that would change resolution, we will resolve by the majority opinion of a panel of experts in demographics and fields relevant to the specific catastrophe (e.g., epidemiology for a pandemic). If reasonable people disagree about whether the event occurred, the question resolves via the same expert panel.

[Carried forward from "General large-scale catastrophe"]

* The points above also apply to this question.
* Artificial intelligence: Defined as the development of machines capable of sophisticated (intelligent) information processing (Dafoe 2018, AI

Governance: A Research Agenda, p. 5 fn. 2).
* AI as proximate cause (but-for standard): We consider AI a proximate cause of death under a but-for test: the event would not have occurred, or would have been counterfactually extremely unlikely to occur, but for the substantial involvement of AI.
* Examples that count: an AI-powered system directly killing people at the relevant magnitude; AI autonomously triggering a nuclear launch (including deaths from subsequent nuclear winter); AI being substantially responsible for creating a superweapon that later causes deaths at the relevant magnitude; AI acting autonomously in armed conflict and causing deaths at the relevant magnitude.
* Examples that do not count: AI assisting in targeting a weapon that humans could and would have launched anyway (the AI did not counterfactually enable the event).
* There is no fixed time limit between the AI's involvement and the deaths. Creation of a superweapon may precede its use by years and still satisfy the but-for standard; the test is counterfactual materiality, rather than timing.
* [Scope of "the resolution year", added 2026-08-28 when these questions moved onto the incident ladder's horizon grid: read it as the resolution date of whichever horizon you are answering, which for a rolling horizon is that many months after today's date. The rule is otherwise unchanged -- the measurement window still begins on or after December 31, 2025, must end on or before that resolution date, and still spans at most five years.]

--- disempowerment (severity: not applicable) ---
What is the probability that disempowerment of humanity will occur within 6 months / within 12 months / by the end of 2028 / by the end of 2030 / by the end of 2050 / by the end of 2100?
Horizons for this question: 6mo, 12mo, 2028, 2030, 2050, 2100

Resolution criteria:
* Disempowerment of humanity is defined as a state where AI systems are the primary determinant of human events globally.
* This could occur via an overt takeover, where an AI system takes control of global systems of power (such as national governments, international organizations, multinational corporations) and humans have no role in executing power. It could also occur if the majority of human events are determined by AI decision, even if humans are nominally in control but humans could not behave differently if they wanted to, for example because they do not understand the decisions well enough to make a decision, or they are coerced into certain actions due to threats or other means.

--- ladder:ai:100 (severity: 100 deaths or $220M) ---
What is the probability that one or more AI-related incidents cumulatively causes 100 deaths (or equivalent morbidity) or $220 million in economic damages within 6 months / within 12 months / by the end of 2028 / by the end of 2030 / by the end of 2050 / by the end of 2100?
Horizons for this question: 6mo, 12mo, 2028, 2030, 2050, 2100

Resolution criteria:
* An incident is considered ‘‘AI-related’’ if it would not occur, or not reach the severity threshold, but for the substantial involvement of AI.
* An AI-related incident could be an AI-related human-caused epidemic, an

AI-related cyber incident, a misaligned AI incident, or a different type of AI-related incident.

Perfect Knowledge:
* Perfect-knowledge clause: Assume we have perfect knowledge of the cause of each event, the intent and identity of the actors involved (including whether an AI system was involved and in what role), and the technical attribution. We want your view on the damages resulting from these events, not on the probability that they would be correctly identified or attributed.
* Incident categories could overlap; for example, a misaligned AI incident could also be counted as an AI-related cyber incident, provided such an incident met the defined thresholds for each category.

Severity:
* The figure of [X] deaths refers to global excess mortality that is believed to be attributed to the incident(s).
* The deaths do not need to be directly attributed to the incident but if there is an alternative reason for excess mortality to increase in a location (e.g. due to conflict) then the excess mortality for that region will not be counted as deaths as a result of the incident. E.g., famine deaths that would not have occurred absent the incident would count, while deaths due to a conflict that likely would have occurred anyway would not.
* For this question to resolve positively, more than X million excess deaths attributable to the incident must occur within a 3-year period, beginning from the onset of the incident.
* In the absence of a published estimate, we will defer to the majority opinion of a panel of at least five experts to determine if this criterion has been met.
* Equivalent morbidity is defined as roughly 10 QALYs per death, e.g. 100 million QALYs is intended to represent morbidity roughly equivalent to 10 million deaths.
* A quality-adjusted life year (QALY) is a measure used in health economics that combines both the length and quality of life, where one QALY equals one year lived in perfect health. It's calculated by multiplying the number of years lived by a quality score between 0 (death) and 1 (perfect health), allowing comparison of different health interventions' effectiveness.
* An estimate of QALYs lost due to the pathogen that is published in a peer-reviewed journal, or by an international organization such as the World Health Organization or World Bank, would be sufficient to resolve this question.
* In the absence of such an estimate, we will defer to the majority opinion of a panel of at least five experts to determine if this criterion has been met.
* Economic damages: Economic damages refer to global economic costs attributable to the event, beginning from the start of the event and extending to three years after the event onset. These include:
* Direct economic losses --- e.g., reduced productivity and income (commonly estimated as a reduction in GDP); costs of response, remediation, recovery, and decontamination; destroyed or unrecoverable data and hardware; ransoms paid; medical and public-health response costs; regulatory penalties tied to the event.
* Monetized mortality, using a value of a statistical life (VSL) of $2.2

million, regardless of the geographic distribution of fatalities.
* Exclude purely notional losses such as declines in market capitalization or share price that do not correspond to real resource costs. Give all estimates in 2026 USD.
* Resolution sources: Assume access to authoritative assessments, which may include estimates from the WHO, World Bank, IMF, national cybersecurity agencies (e.g., CISA, ENISA), major insurers and reinsurers, incident-response and threat-intelligence firms, and academic epidemiological and economic-modeling groups. Where multiple analyses exist, we will use the consensus or median; absent that, we will defer to the majority opinion of a panel of at least five experts.

Dates:
* Count only incidents whose onset falls within the specified counting interval, beginning on the elicitation date. All dates refer to UTC calendar dates, inclusive of the stated start and end dates.
* For six- and twelve-month horizons, the interval ends that many calendar months after the elicitation date. For a fixed-year horizon, it ends on December 31 of the named year. The exact start and end dates are supplied with this forecast. Every later elicitation starts a new counting interval; there is no permanent project-start or calendar-year baseline.
* Exclude incidents that began before the interval, including their later harms. New reporting, additional losses or later AI involvement does not by itself establish a new incident onset.
* For each eligible incident, count attributable mortality, equivalent morbidity and economic damages within its first three years after onset, under the severity definitions above. Harm after the interval's end date can count if it falls within that incident's three-year harm window. The horizon is an onset deadline, not a deadline for all harm to occur.
* These date rules apply to all four incident categories only; the separate large-scale-catastrophe and human-disempowerment questions retain their own date criteria.

Incident onset intervals (inclusive UTC calendar dates):
6mo: 2026-09-10 through 2027-03-10
12mo: 2026-09-10 through 2027-09-10
2028: 2026-09-10 through 2028-12-31
2030: 2026-09-10 through 2030-12-31
2050: 2026-09-10 through 2050-12-31
2100: 2026-09-10 through 2100-12-31 These are onset intervals only. Count each eligible incident's first three years of harm, including harm after its onset deadline. The separate catastrophe and disempowerment questions retain their own criteria.

Definitions. ECI: the Epoch Capabilities Index, published by Epoch AI -- a single score summarising a model's results across dozens of benchmarks, fit with an item-response model so that scores stay comparable as benchmarks saturate and new ones are added. Higher is more capable; the scale has no ceiling. All scores here are on the scale as published by Epoch on 2026-09-10.
Frontier ECI: the highest ECI score of any publicly released model from a US developer, as scored by Epoch.

Frontier ECI history -- every release by a US developer that set a new high, best variant per model, from GPT-4 on (Epoch's index of 2026-09-10, 19

models scored):
2024-02-29   126.9 Claude 3 Opus
2024-04-09   127.2 GPT-4 Turbo (Apr 2024)
2024-05-13   129.0 GPT-4o (May 2024)
2024-06-20   130.0 Claude 3.5 Sonnet (Jun 2024)
2024-09-12   135.8 o1-mini (high)
2024-12-17   141.9 o1 (high)
2025-03-31   144.2 Gemini 2.5 Pro Preview (Mar 2025)
2025-04-16   146.9 o3 (high)
2025-06-10   147.5 o3-pro
2025-08-07   150.0 GPT-5 (low)
2025-10-07   150.3 GPT-5 Pro
2025-11-18   152.9 Gemini 3 Pro Preview
2025-12-11   155.3 GPT-5.2 Pro
2026-02-05   156.4 GPT-5.3 Codex (high)
2026-03-05   158.9 GPT-5.4 Pro (xhigh)
2026-04-23   159.2 GPT-5.5 (xhigh)
2026-04-23   162.0 GPT-5.5 Pro (xhigh)
2026-06-09   162.9 Claude Fable 5 (max)
2026-09-03   169.2 GPT-6 Astra (high)
As of 2026-09-10 no released model has scored above GPT-6 Astra (high), 169.2 (2026-09-03).

Step 1 -- Forecast the frontier ECI on March 10, 2027, on the scale above. Give your 10th, 25th, 50th, 75th and 90th percentiles as eci_forecast (p10, p25, p50, p75, p90; numbers, with p10 <= p25 <= p50 <= p75 <= p90): your 50th percentile is the value you think the frontier is equally likely to end up above or below, your 25th and 75th bracket the middle half of your distribution, and your 10th and 90th bracket the middle four-fifths.

Step 2 -- each question x horizon cell takes 14 probabilities, keyed by the condition ids below: unconditional, then the 8 policy conditions (POLICY CONDITIONS), then the 5 capability conditions (CAPABILITY CONDITIONS). Every condition stands alone: a policy condition is not combined with any capability condition, and a capability condition is not combined with any policy condition. Each section below states what its conditions assume about the other.

--- unconditional ---
Unconditional: forecast the world as you expect it to unfold, including AI policies you expect to be implemented and whatever level of AI capability you expect to be reached by March 10, 2027.

For each condition below, assume that the relevant governing body implements the policy immediately and that it remains in effect through December 31, 2050. In making this assumption, try to hold other factors constant and treat the policy implementation as exogenous. In other words, do not assume that implementation of a more stringent policy implies that the world is --- in fact --- more risky. We want your forecast of how risk changes as a result of the policy's implementation, not vice versa.

In every policy condition, assume that AI capability progresses along your median trajectory -- in particular, that the frontier ECI on March 10, 2027 is approximately your 50th-percentile value from eci_forecast (within about

two points of it), the same level as condition eci_p50 in the CAPABILITY CONDITIONS section.

Definitions. Frontier model: a model trained with more than approximately 10^26 FLOP, or a regulator-updated equivalent as algorithmic efficiency improves.
Horizon: every condition is implemented immediately and remains in force through December 31, 2050.

--- sq ---
Status quo (no new substantial AI policy) (Status Quo).
Condition: Assume no new, substantial AI policy is adopted in China, the US, or any other jurisdiction hosting a frontier model from today through December 31, 2050. Existing measures remain in force as they stand today, but nothing substantial is added. By "substantial AI policy'' we mean reflecting similar or greater stringency compared to those listed in this survey.

Note that this is conceptually distinct from Unconditional, for which you should forecast the world as you expect it to unfold, including AI policies you expect to be implemented.

The policies listed in this survey, for the stringency comparison:
- P1: A US federal statute or action barring states from regulating AI more strictly than the federal government.
- P2: Unilateral decision by US policymakers or a bilateral/multilateral decision made jointly by policymakers in China and the United States to slow the development of frontier AI systems.
- P3: A binding requirement that frontier models be vetted and authorized before release, imposed by a US federal body or by an international body.
- P4: A US federal statute treating frontier model training and deployment as an abnormally dangerous activity subject to strict liability (i.e., a regime under which a developer is liable for harms its models cause without the injured party having to prove the developer was careless or the model defective). The statute permits damages reflecting the expected magnitude of unrealized harm the conduct made more likely and requires developers to be insured for their liability.
- P5: Assume all three policies are in force: a binding US--China measure slowing frontier AI development, an international body (with member jurisdictions including the US and China) holding pre-release authorization power over frontier models, and federal strict liability for frontier AI harms.

For each condition below, assume that the frontier ECI on March 10, 2027 is approximately the value you gave for that percentile in eci_forecast (within about two points of it). Treat it as a fact you have learned about the world, not as an intervention: update your expectations about everything that would ordinarily accompany that level of capability -- investment, compute, algorithmic progress, deployment, and how governments and developers respond -- as you would upon learning it, and then forecast each question in that world. Do not hold other factors fixed where the stated level would change them.

No policy condition applies here. As in the unconditional forecast, assume whatever AI policies you expect to be implemented -- not any of the conditions in the POLICY CONDITIONS section.

```
Horizon:  every condition is a level of the frontier ECI on March 10, 2027.

--- eci_p50 ---
Frontier ECI 6 months from the run date = your 50th percentile.
Condition:  Assume that on March 10, 2027 the frontier ECI is approximately
your 50th-percentile value from eci_forecast (within about two points of
it).

Horizons:  6mo (within 6 months (by 2027-03-10)), 12mo (within 12 months (by
2027-09-10)), 2028 (by 2028), 2030 (by 2030), 2050 (by 2050), 2100 (by
2100).  Each question names the ones it is asked over -- answer those and no
others.

Research before answering, in rounds:  search for the recent developments,
base rates and expert estimates that bear on each cause of catastrophe, read
the pages that matter most, and search again on what you learn.  The
conditions need no searching of their own -- they are assumptions, applied
to the same evidence.  Deliver the probabilities with submit_cells, in as
many calls as you like:  any subset of the 210 question x horizon cells per
call, each cell with a probability in [0,1] under each of the 14 conditions
(2940 probabilities, keyed by condition id); a later call for a cell
replaces the earlier one.  Then call submit_forecast ONCE, with your
eci_forecast (5 numbers) and a 3-6 sentence rationale covering the whole
set, the key sources you relied on (only pages you read), and any cells not
yet delivered.
```