Conditional Regulation of Frontier AI with Automated Insider Forecasts
Jason Abaluck
June 2026

BASIC IDEA: use insider AI-assisted forecasts of bad event probabilities as a regulatory trigger, rather than waiting for dangerous capabilities to be revealed in red teaming or post-deployment

PROPOSAL:

- An external or government agency develops a suite of forecasting models, built on frontier models. These are “evaluators”
- Evaluators have read-only access to code, training runs, and research notes at every frontier lab
- Each evaluator independently forecasts, conditional on current/planned R&D continuing, the probability of specific dangerous events: e.g. P(biodisaster|deployment) at 1, 3, 6, and 12-month horizons or P(recursive self-improvement|deployment), as well as forecasts validation events expected to occur more frequently
- Output: a dashboard of probabilities of bad events over time
- This probability should be a tripwire for regulatory oversight
- This tripwire builds in natural *pause and unpause* conditions -- you unpause when the forecasted probability of bad events given resumption falls for a fixed period
- Performance must be validated relative to human forecasters before deployment

RATIONALE:

Frontier AI creates various catastrophic risks, ranging from enabling biodisasters to recursive self-improvement leading to human extinction. Companies themselves don’t fully internalize the downside risk to society, so external regulation is needed. Regulators need near real-time information to accurately assess risk; only models with white-box access can achieve this. A suite of models, each with different provenance, is needed to reduce risk of model collusion. Regulation needs to kick in before dangerous capabilities are demonstrated. Forecasts allow this to happen. At present, proven human forecasters outperform the best models given similar information, but model forecasts are improving as capabilities increase; model forecasts should be supplemented with lower-frequency superforecaster predictions until model parity with superforecasters is validated.

FULL PROPOSAL:

Why are conditional forecasts needed?

Many existing regulatory proposals such as Karnofsky's "If-Then Commitments for AI Risk Reduction" (Carnegie, Sept 2024) and "Tripwire Capabilities" (Carnegie, Dec 2024) tie mitigations to achieved capabilities. But achieved capabilities may be poor proxies for danger, or arrive only after a dangerous transition in unmeasured capabilities has already occurred. Several frameworks call for forward-looking risk estimates and safeguards prior to model deployment (UK Government 2024; OpenAI 2025; METR 2025). This proposal provides a quantitative framework for achieving this. Rather than banning AI development entirely and risk foregoing its benefits, we want to set appropriate speed limits (Koh & Sanguanmoo 2026), and conditional forecasts provide the right space in which to set speed limits.

Conditional forecasts (if accurate) are a key input to determine when costly action to mitigate risk passes a cost-benefit test. To be fully specified, the conditional forecasts being made need the same level of specificity as typical contracts on Polymarket or Kalshi. Examples of fully specified conditional probabilities are given in Technical Appendix A.

Can Models Forecast Catastrophic Risk?

Models are not as good as the best humans, but models are improving (Metaculus 2025), and the advantage of an automated model-based forecast is that it can produce a high-frequency panel to gauge changes in risk in response to new information. Model-based forecasts can also consider a broader range of events than public prediction markets, which require liquidity on each contract to make accurate predictions (Wolfers & Zitzewitz 2004). The subsequent two figures provide new evidence that model forecasts are continuing to improve.

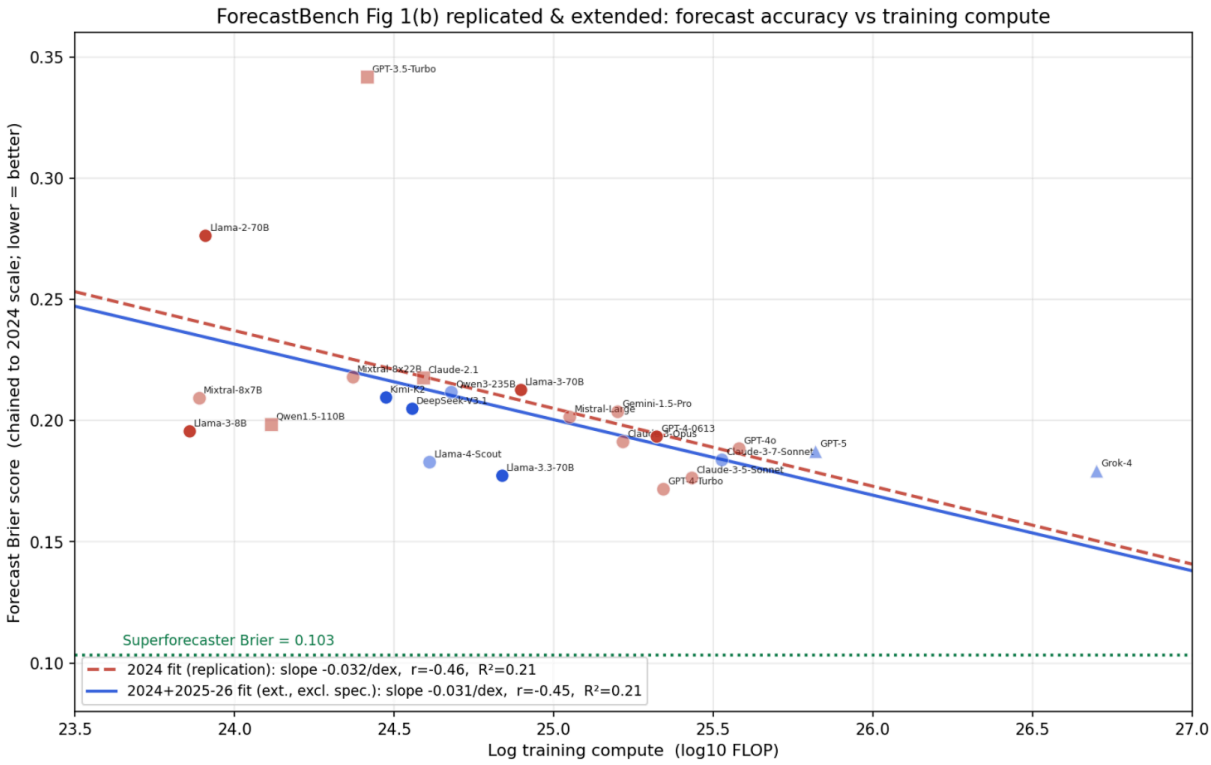

Figure 1: Forecast accuracy vs. training compute (Brier; lower = better). It replicates and extends the result from Karger et al. 2024 that models pre-trained with more compute have better forecast accuracy. For more recent models, relevant parameters to compute training compute are not disclosed (and pre-training compute becomes a less relevant measure for reasoning models with substantial test-time compute).

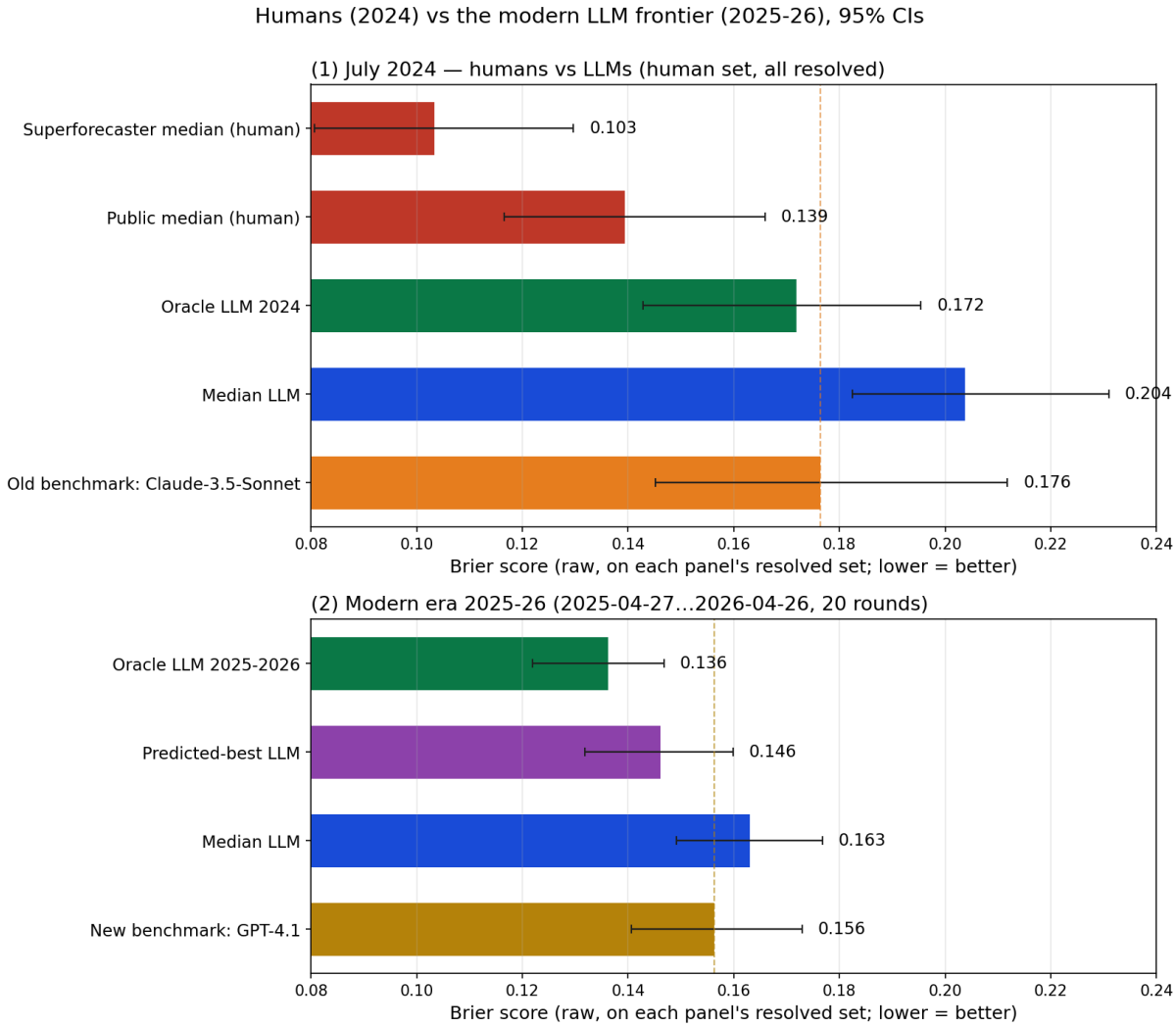

Figure 2: Forecasting accuracy of humans vs. the best LLMs, July 2024 (top) and 2025–26 (bottom). The top panel shows the comparison between superforecasters, median humans, and LLMs in the original Karger et al. 2024 analysis. The bottom panel uses ForecastBench to extend these results, showing the best recent models (human forecasts are not available in the "Modern Era"). No model is present for the entirety of both periods; GPT 4.1 is present for the entirety of the modern era and Claude 3.5 Sonnet is present for the entire initial period; GPT 4.1 is a stable ~0.015 Brier better for the periods where they overlap, suggesting that the prior and modern period forecasts are of roughly equal difficulty (perhaps modern questions are 0.005 Brier easier). Oracle LLM shows the average Brier score of the single best LLM in each period, while predicted-best LLM shows the average Brier score of the LLM predicted to be best using only pre-period information (eliminating the "winner's curse" bias from selecting a winning model that happened to be lucky).[1]

The bottom-line of this analysis: Brier scores of the best models have continued to improve:

[1] Replication instructions are available at: https://github.com/Jabaluck/forecastbench-replication

testing the predicted-best in 2025-2026 vs. the best persistent models from the 2024 analysis gives two-sided p = 0.013. The best LLMs now appear comparable to the median of public human forecasts, although still below the level of proven superforecasters.

The proposal detailed here differs from the ForecastBench in that models would have an informational advantage over the best humans due to the scope of the available insider information. That said, human-assisted forecasts by validated superforecasters utilizing model forecasts as inputs may well outperform even inside-information assisted model forecasts for the time-being.

Several steps require validation before we can conclude that models can forecast catastrophic risk:

1) Can models with inside information outperform publicly available forecasts of intermediate variables? Such variables could include dates of flagship model releases, performance of unreleased models on future benchmarks, or forecasts of changes in political economic variables (such as company valuations) following model release.
2) Can models correctly forecast probabilities of intermediate phenomena related to catastrophic risk? These include:
    a) Ability of models to construct other models hitting various benchmarks (a precursor to recursive self-improvement). For example, Anthropic reports that Claude Mythos Preview, given code that trains a small neural network and asked to make training faster, achieved roughly a 52x speedup, compared to 4x for a skilled human (Anthropic 2026). This exercise could be repeated for more sophisticated models, (e.g. earlier Opus or Sonnet models). What speedups could be achieved by Mythos and future models in training existing models, and are these speedups forecastable?
    b) Biosecurity failures short of disaster, other biorisk benchmarks; e.g., performance on biology and wet-lab troubleshooting benchmarks (such as virology and protocol-debugging tasks), or model-assisted removal of wet-lab bottlenecks
    c) Persuasion benchmarks, such as success at persuading people to take costly actions in the physical world with a fixed budget (e.g. induce people to buy a consumer product with a fixed budget, exercise today, or email someone asking for an intro)
3) In cases 1) and 2), can models with insider information outperform superforecaster or prediction market forecasts of the same quantities?

<u>Can This Be Combined with Prediction Markets?</u>

Yes, and ideally it should be. Models should use as one input to their forecasts the prices available on public prediction markets for related questions. Labs could also run internal prediction markets where employees were given seed funding which they were required to use to trade on those markets to generate sufficient liquidity (Fradkin et al. 2026).

Will Models Collude?

A potential risk as models become sophisticated is that models will collude with evaluators, e.g., offering payoffs for the evaluator to give misleading forecasts (Yudkowsky 2002; Greenblatt et al. 2023). Several steps potentially reduce this risk:

- Use a variety of evaluator models with different provenance, hidden from labs
- Blind labs to which evaluator models are used
- Evaluators should have short-leash alerts for any detected communication that set off harsh regulatory interventions

This proposal is not a substitute for other safety measures, such as chain of thought monitoring and red teaming among others.

Can Lab Proprietary Information Be Protected?

If you give a Claude instance access to all of OpenAI's code, there is a serious concern that Anthropic would now be able to reverse engineer that code. So evaluator access cannot mean sending lab materials through ordinary commercial APIs.

Evaluator models would need to be run inside highly secure environments with no outbound network access and many other safeguards in place. The broad rule should be that data is sent to the evaluators, but the evaluators output nothing except the forecast probabilities, and sealed explanatory records for regulatory review.

Who Should Oversee This?

Labs themselves have too little incentive to reduce existential risk. A program like this could be piloted by labs, but ultimately, external regulations are necessary.

The program should be administered either by a government agency or a third party agency contracted by the government to run evaluators, with regulatory triggers tied to evaluator probabilities. Labs above a compute or capability threshold would be required to maintain auditable records, provide evaluator access, and receive authorization for training runs and deployments above specified thresholds of forecasted risk.

What Regulations Should be Triggered?

The most natural regulation is a pause on further frontier model development until conditional disaster probabilities recede. Following this pause, an extensive human audit could be conducted to assess safety, with activity resuming only if forecasted conditional disaster probabilities fall. Importantly, this proposal builds in a natural criterion for *unpausing* -- when forecasted conditional disaster probability given unpausing recedes.

How Will Labs Respond to This Regulation?

Labs will seek to avoid loss of power. If labs respond to this regulation by changing behavior to lower catastrophic risk --- e.g. additional caution regarding recursive self-improvement --- then this is precisely the intended response.

An alternative possibility is that labs, or autonomous models, respond by hiding information from evaluators in order to avoid a shutdown. Willful concealment that materially impairs catastrophic-risk oversight should trigger escalating consequences, including major civil penalties, suspension or revocation of authorization to train or deploy frontier models, and criminal liability under applicable false-statement, obstruction, and public-safety statutes, with enhanced penalties where concealment foreseeably creates catastrophic public risk. Regular protocols and audits should verify the completeness of the information fed to evaluators.

A potentially more concerning threat is evasion by models. An important alignment question beyond the scope of this note is how to prevent frontier models from hiding code from evaluator models. It is unlikely this can be entirely prevented, but sophisticated evaluator models should also take this possibility into account in their forecasts, and models found to be deliberately hiding code from evaluators should be subject to increased oversight or constraints.

How Should Risks be Aggregated and What Probability Threshold Should be Used?

If the forecast dashboard tracks multiple risks, a pre-specified rule should be used to aggregate these risks into a bottom-line headline (which might, for example, be based on the maximum of the risks of various catastrophic events).

The threshold for shutdown is difficult to determine, as even a modest level of existential risk could justify a shutdown. Clearly, very high levels of short-term risk are unacceptable. For example, if the risk of an extinction event within the next year exceeded 10%, then immediate regulatory action is necessary. Determining at which thresholds regulatory action should kick in is an important topic for separate analysis.

In principle, forecasts of *counterfactuals* are more relevant for decision-making than the conditional forecasts above. We want to know not only the probability of various risks, but how that probability changes if various regulatory solutions are attempted. However, forecasts of counterfactuals will likely be highly dependent on the details of the regulatory solution, so forecasts of conditional risks are more informative that some action is needed, although they do not reveal the right action.

Acknowledgements
Thanks to Tom Cunningham, Aly Murray, Chris Painter, Phil Trammell and Cheryl Wu for helpful discussions.

Bibliography

Anthropic. (2026). Claude Mythos Preview System Card.
https://www.anthropic.com/claude-mythos-preview-system-card
Fradkin, A., Jabarian, B., & Koh, A. (2026). We need well-capitalized prediction markets for AI impacts. Justified Posteriors (Substack).
https://empiricrafting.substack.com/p/we-need-well-capitalized-prediction
Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942. https://arxiv.org/abs/2312.06942
Halawi, D., Zhang, F., Yueh-Han, C., & Steinhardt, J. (2024). Approaching Human-Level Forecasting with Language Models. NeurIPS 2024; arXiv:2402.18563.
https://arxiv.org/abs/2402.18563
Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., & Tetlock, P. E. (2024). ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. arXiv:2409.19839.
https://arxiv.org/abs/2409.19839
Koh, A., & Sanguanmoo, S. (2026). Technology Speed Limits. arXiv:2606.01424.
https://arxiv.org/abs/2606.01424
Metaculus. (2025). AI Forecasting Benchmark (FutureEval): Q1 and Q2 2025 Results.
https://www.metaculus.com/aib/2025/q2/
Wolfers, J., & Zitzewitz, E. (2004). Prediction Markets. Journal of Economic Perspectives, 18(2), 107–126. https://doi.org/10.1257/0895330041371321
Yudkowsky, E. (2002). The AI-Box Experiment. https://www.yudkowsky.net/singularity/aibox
METR. (2025). Common Elements of Frontier AI Safety Policies.
https://metr.org/common-elements-of-frontier-ai-safety-policies
OpenAI. (2025). Preparedness Framework (Version 2).
https://openai.com/index/updating-our-preparedness-framework/
UK Government. (2024). Frontier AI Safety Commitments, AI Seoul Summit 2024.
https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

TECHNICAL APPENDIX 1:

Formal specification of conditional probabilities

Because catastrophic events are rare and largely unprecedented, the same template is applied to higher-frequency validation and intermediate events: model release dates, unreleased-model benchmark performance, security incidents short of disaster, autonomous-replication indicators, and persuasion-benchmark results, which resolve often enough to test calibration and the evaluators' insider-information advantage.Each conditional forecast must be specified precisely enough to be unambiguously resolvable, at the level of a Polymarket or Kalshi contract. The table below lists the components that every forecast must fix, and shows how an example forecasts of an AI-enabled biological incident applies in each case.

Component	**Example: AI-enabled biodisaster (OpenAI GPT-7)**
(i) Conditioning event (currently authorized or requested trajectory)	OpenAI deploys GPT-7 with general API access before June 1, 2028, without implementing enhanced biosecurity safeguards (e.g., bio-uplift refusal training, know-your-customer screening for high-risk biology queries, and tiered access controls)
(ii) Outcome event and severity threshold	An AI-assisted biological incident causing at least 1 million human fatalities
(iii) Resolution horizon	By June 1, 2029
(iv) Geographic / population scope	Anywhere worldwide; human fatalities
(v) Causal role of the model	Access to GPT-7 is a but-for or substantial contributing factor

TECHNICAL APPENDIX 2:

Replication details for figures

All ForecastBench quantities are computed from publicly available data at forecastbench.org. For each model and biweekly round we compute a difficulty-adjusted Brier score, 0.5·mean(adjusted dataset Brier) + 0.5·mean(adjusted market Brier), following Karger et al. (2024), where each question's raw Brier score is adjusted by its published question fixed effect; across the 68 matchable models this reproduces ForecastBench's published per-model leaderboard closely.

Figure 1 (training compute). Replicating Karger et al. (2024), training-compute estimates are taken from Epoch AI; the relationship between log10 training compute and the Brier score is fit by ordinary least squares. Estimates are unavailable for most recent closed frontier models, and pre-training compute is a weaker proxy for reasoning models with substantial test-time compute, so the extension covers only the subset of models with disclosed or estimated compute.

Figure 2, top panel (July 2024). Superforecaster and public-median Brier scores are computed on the 2024-07-21 human question set; this is the only set humans forecasted. Using all resolved questions; each LLM is scored on the identical questions, so the comparison is apples-to-apples and question difficulty cancels. The single-best ('oracle') and median LLMs are taken over the models present in that round, and the old benchmark is Claude-3.5-Sonnet.

Figure 2, bottom panel (2025–26). Because the frontier models differ across the two eras, the panels are linked through benchmark models present in both. Each panel plots the benchmark with contemporaneous data in that era: the old benchmark (Claude-3.5-Sonnet, orange) in July 2024 and the new benchmark (GPT-4.1, mustard) in 2025–26. These two are both fully-present in their own era while also overlapping: across their six overlap rounds (April–August 2025) GPT-4.1 scores a stable 0.015 Brier better than Claude-3.5-Sonnet (largest round-level deviation 0.026). This near-constant offset places both panels on a common scale and anchors the cross-era superforecaster-versus-modern-oracle comparison. The oracle LLM is the per-round best; the predicted-best LLM is selected in each round using prior-round performance only, which removes the winner's-curse bias from picking an ex-post winner (see the replication package for details of this prediction). Confidence intervals are two-level bootstraps, resampling rounds, then questions within rounds, and the benchmark-equivalence offset is bootstrapped over the overlap rounds.