On Forecasting Lags in AI Risk Evaluation

Paolo Bova

Teesside University
paolobova@proton.me

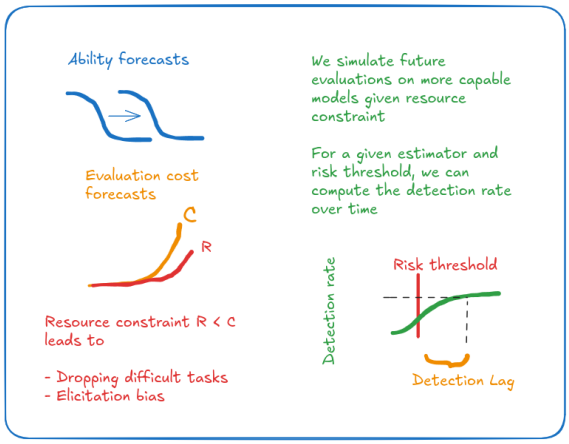

Figure 1: A detection lag model for AI Risk evalutions. Using trend data on evaluation costs and AI ability, simulates evaluations under plausible resource constraints. We can then compute lags in spotting a crossing of a set threshold for a given estimator computed on the simulated data.

Introduction

Aritifical Intelligence (AI) risk evaluations are key to building safety cases for new AI systems (Shevlane et al. 2023; Buhl et al. 2024; Goemans et al. 2024). Without them, we would have little understanding of the ways in which AI systems could cause harm or how quickly they are gaining competencies or propensities to engage in harmful behaviour. These evaluations are often led by AI companies, governments, or public organisations focused on AI risk evaluation and cover a number of risks, including risks from enabling the development of weapons, risks from cyberattacks, risks from model autonomy, deception, and persuasion (Google DeepMind 2024; Phuong et al. 2024; Kinniment et al. 2024; Benton et al. 2024).

Investment in high-quality AI risk evaluations appears to be essential to raise the bar for the safe development of AI systems (Bova, Di Stefano, and Han 2024). However, we currently lack research into forecasting how the *size* of this investment affects our ability to provide early warnings for new AI risks.

Model

We contribute a framework for modelling lags in AI risk evaluation, building on our prior work (Bova, Stefano, and Han 2024). In summary, the model uses current data on trends in AI ability and the costs of evaluations to predict the performance of future AI systems in more difficult evaluations that may face budget constraints.

First, we forecast model ability over time. We aim to specify the full shape of the success rate curve (i.e. characteristic curve) at each time step. Using the data and replicating the logistic regressions from Kwa et al. (2025), we observe that while the time horizon estimates follow an exponential trend, the slope of the logistic curves are quite variable. For the sake of simplicity, we fixed this slope to be the average of the slope fits for that data, and used their time horizon trend to predict the model ability every 3 months from 2025 to 2030. Their trend predicts AI systems double their performance roughly every 7 months.

Second, using the same data and after controlling for the typically lower costs of failed tasks, we forecast the costs of evaluation tasks as our measure of difficulty rises. To do so, we performed a regression of $log_2(\textit{generation cost})$ against task difficulty $log_2(\textit{human seconds})$ for each AI model in the data. We then took the geometric mean of the predicted doubling rates for the generation cost of a task, which we will now denote by d, finding that $d \approx 1$ unit of difficulty. So, a doubling in the number of seconds it would take a human to complete the task typically led to a doubling in cost. This result matches the intuition that under the inference compute paradigm, more tokens can be used to answer problems that require longer feats of reasoning.

Our third step is to specify what a sufficient gold standard evaluation for AI risk involves. We define a gold standard evaluation at time t, as an evaluation that includes tasks that cover a sufficiently large region of the characteristic curve at time t. Specifically, this evaluation window must be wide enough to cover tasks with as low as 10% success rate to tasks with 90% chance. If we set $\bar{x}$ as the threshold 50%, then the evaluation window is $[b = \bar{x} - \frac{\delta}{\textit{slope}}, u = \bar{x} + \frac{\delta}{\textit{slope}}]$. From our cost forecast, it is reasonable to model

the cost as a function of difficulty x as follows: $c(x) = 2^{\frac{x}{d}}$, normalised so that $c(0) = 1$. We can then compute the total cost of the gold standard evaluation, $C = nE[c]$, where $n = r(u-b)$ is the number of tasks we sample for the evaluation, which we assume to be equal to r repeats per difficulty unit. $E[c]$ denotes the expected cost of each task, assuming uniform sampling. We can integrate the cost function over the evaluation window to find $E[c] = (2^{\frac{u}{d}} - 2^{\frac{b}{d}}) \cdot \frac{d}{u-b} \frac{1}{\log 2}$.

We now introduce a budget constraint, $R \leq C$, typically as a fraction of C. Assume that, when facing this constraint, we drop the most difficult tasks first. In that case, the low end of the evaluation window, b, is unchanged. Setting our equation for the total cost equal to R, we can solve the new upper end as $u_{new} = d \cdot \log_2(\frac{R \log 2}{rd} + 2^{\frac{b}{d}})$.

We then run Monte Carlo simulations where in each run and at each time step t, an evaluation is created by sampling tasks from this new evaluation window. The tested AI system then attempts each task, with success rates following the predicted characteristic curve at time t.

We compare the effectiveness of different estimators for AI capability. Our first estimator is the same logistic estimator used to estimate the time horizons of the AI systems in Kwa et al. (2025). We choose a time horizon of 2^{20} seconds as our capability threshold, i.e. roughly 2 weeks.

Our second estimator intends to mimic and extend a business-as-usual approach to tracking AI risk. In most works, performance is measured as a percentage of solved tasks in a benchmark. Even in Phuong et al. (2024) where critical capability thresholds are specified, they are specified as a proportion of the selected tasks being achieved by the AI system. On it's own, a simple success rate doesn't tell us much, so we argue that there is an implicit measure of the severity of the selected tasks that is being considered when setting such a threshold. When this is extended to multiple levels, a natural description of such a measure is as a weighted sum of the success rates in each level. This definition is vital as when resource constraints reduce the difficulty of tasks selected for the benchmark, this leads to the benchmark being considered as capturing a lower threat level than before. For consistency, we set the threshold to be the weighted score achieved by a model that is at the logistic estimator's threshold. For the weighting function, $f(x) = x$, this will be roughly 62.5.

Finally, from our Monte Carlo simulations we can compute the distribution of each estimator at each time step, $\hat{y}_t$. We can then compute the detection likelihood for a set threshold, y^*, as $Pr(\hat{y}_t \geq y^*)$. Treating detection itself as a random variable, we compute the detection lag as the median extra time Δ_t before detection occurs conditional on detection happening at all.

Results

A direct comparison of the two detection rate plots in Figures 2 and 3 shows that the logistic estimator performs well even under large budget constraints. The steep slope around the threshold shows that the detection rate spikes exactly when needed. On the other hand, the business-as-usual (weighted score) estimator performs well only with close to a full budget; otherwise, the detection rate curve shifts to the right, introducing a detection lag. It also becomes increasingly likely that no detection will occur at all over the entire time period.

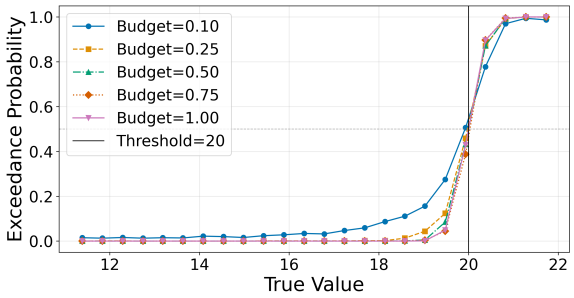

Figure 2: Detection rate as AI systems become more capable for different budget constraints. Here we use the logistic estimator and set a capability threshold of 2^{20} seconds.

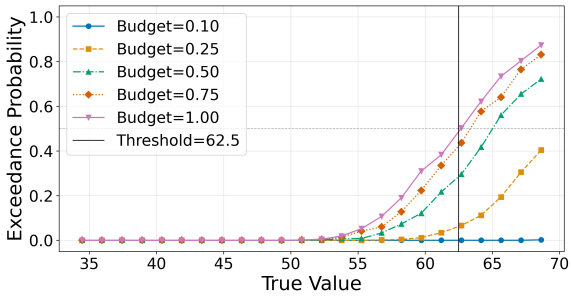

Figure 3: Detection rate as AI systems become more capable for different budget constraints. Here we use the business-as-usual (weighted score) estimator and set a threshold of 62.5 which is roughly equivalent to the weighted sum a model would get if it achieved the 2^{20} capability threshold. The weighting function is $f(x) = x$.

We find in our simulations that a business-as-usual approach to tracking AI risks is vulnerable to substantial detection lags when facing budget constraints. These detection lags were not present when using the logistic estimator proposed by METR. We advise greater adoption of this logistic estimator where appropriate, to help mitigate against such lags (Kwa et al. 2025). This requires more careful design of AI risk evaluations to track the difficulty of different tasks. Useful works in this direction include (Kwa et al. 2025; Zhou et al. 2025).

Inspired by these initial findings, we expect that this model may also prove useful for identifying gaps in funding for future AI risk evaluations. To be most useful, we envisage three main ways to extend and generalise our framework:

- Measure how elicitation bias varies with research time.
- Extend these forecasting methods to multidimensional scores of risk (Zhou et al. 2025).
- Find additional proxies for evaluation costs (e.g. research time, pre-deployment test time, staff budgets).

Acknowledgments

I am grateful to my supervisors, The Anh Han and Alessandro Di Stefano, for their constant support in the production of this work.

References

Benton, J.; Wagner, M.; Christiansen, E.; Anil, C.; Perez, E.; et al. 2024. Sabotage Evaluations for Frontier Models. *ArXiv*.

Bova, P.; Di Stefano, A.; and Han, T. A. 2024. Both eyes open: Vigilant Incentives help auditors improve AI safety. *Journal of Physics: Complexity*, 5(2): 025009.

Bova, P.; Stefano, A. D.; and Han, T. A. 2024. Quantifying detection rates for dangerous capabilities: a theoretical model of dangerous capability evaluations. arXiv:2412.15433.

Buhl, M. D.; Sett, G.; Koessler, L.; Schuett, J.; and Anderljung, M. 2024. Safety Cases for Frontier AI. arXiv:2410.21572.

Goemans, A.; Buhl, M. D.; Schuett, J.; Korbak, T.; Wang, J.; et al. 2024. Safety Case Template for Frontier AI: A Cyber Inability Argument. arXiv:2411.08088.

Google DeepMind. 2024. Introducing the Frontier Safety Framework.

Kinniment, M.; Sato, L. J. K.; Du, H.; Goodrich, B.; Hasin, M.; et al. 2024. Evaluating Language-Model Agents on Realistic Autonomous Tasks. arXiv:2312.11671.

Kwa, T.; West, B.; Becker, J.; Deng, A.; Garcia, K.; Hasin, M.; Jawhar, S.; Kinniment, M.; Rush, N.; Arx, S. V.; Bloom, R.; Broadley, T.; Du, H.; Goodrich, B.; Jurkovic, N.; Miles, L. H.; Nix, S.; Lin, T.; Parikh, N.; Rein, D.; Sato, L. J. K.; Wijk, H.; Ziegler, D. M.; Barnes, E.; and Chan, L. 2025. Measuring AI Ability to Complete Long Tasks. arXiv:2503.14499.

Phuong, M.; Aitchison, M.; Catt, E.; Cogan, S.; Kaskasoli, A.; et al. 2024. Evaluating Frontier Models for Dangerous Capabilities. arXiv:2403.13793.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; et al. 2023. Model Evaluation for Extreme Risks. arxiv:2305.15324.

Zhou, L.; Pacchiardi, L.; Martínez-Plumed, F.; Collins, K. M.; Moros-Daval, Y.; Zhang, S.; Zhao, Q.; Huang, Y.; Sun, L.; Prunty, J. E.; Li, Z.; Sánchez-García, P.; Chen, K. J.; Casares, P. A. M.; Zu, J.; Burden, J.; Mehrbakhsh, B.; Stillwell, D.; Cebrian, M.; Wang, J.; Henderson, P.; Wu, S. T.; Kyllonen, P. C.; Cheke, L.; Xie, X.; and Hernández-Orallo, J. 2025. General Scales Unlock AI Evaluation with Explanatory and Predictive Power. arXiv:2503.06378.