
A Multi-Domain Stochastic Framework for Forecasting Catastrophic Risk from Artificial Intelligence Development Through 2030¹

Ibrahim El-chami, Ph.D.

Artificial Intelligence Research and Development Studios
Electrical and Computer Engineering, The University of British
Columbia*
ibrahim@ai-rd.ca

With
Apart Research

Abstract

We present a novel integrated forecasting framework that synthesizes five distinct statistical methodologies: Solow-Romer endogenous growth theory for compute scaling, Generalized Extreme Value theory for capability forecasting, Nash equilibrium game theory for geopolitical dynamics, Computable General Equilibrium economics, and Hawkes self-exciting point processes for catastrophic risk assessment. The assessment is to project AI-related systemic risks through 2030. Employing Monte Carlo simulation with 500 trajectories per scenario, we evaluate four policy intervention frameworks: baseline (no coordination), global coordination, Western alliance (excluding China), and unilateral US regulation.

Our findings indicate that under baseline assumptions, the cumulative probability of at least one catastrophic AI incident by 2030 reaches 26% (90% CI: 18-35%). Global coordination reduces this to 12% (54% relative reduction), while paradoxically, partial coalitions increase risk to 30% due to arms race/acceleration dynamics. We identify a critical intervention window between 2025-2027, after which autonomous systems reduce governance effectiveness by approximately 60%. The model projects automation rates reaching 38% by 2030 under baseline conditions, exceeding the 30% economic disruption threshold with corresponding unemployment impacts of 11%. All 147 model parameters derive

¹ Research conducted at the Apart Research - AI Forecasting Hackathon, 2025
<https://apartresearch.com/sprints/the-ai-forecasting-hackathon-2025-10-31-to-2025-11-02>

from peer-reviewed literature, authoritative government reports, or verified industry data, ensuring empirical grounding. This work contributes a methodologically rigorous, multi-domain approach to AI risk assessment that explicitly models feedback loops and tail risks while maintaining full transparency in data provenance and uncertainty quantification.

Keywords: AI safety, catastrophic risk, extreme value theory, game theory, Monte Carlo simulation, policy analysis, technology forecasting, self-exciting processes

1. Introduction

The rapid advancement of artificial intelligence systems presents both unprecedented economic opportunities and novel existential risks (Bostrom, 2014; Amodei et al., 2016). Recent developments, including the emergence of large language models with capabilities approaching human-level performance on complex reasoning tasks (OpenAI, 2023; Anthropic, 2024) and breakthrough cost-efficient training methods (DeepSeek-AI, 2024), have accelerated timelines for transformative AI. Concurrently, geopolitical tensions between major AI powers have intensified competition dynamics (Allen, 2023; CSIS, 2024), while rapid automation threatens labor market stability (Acemoglu & Restrepo, 2020; Brynjolfsson et al., 2024).

Despite growing recognition of these interconnected risks, existing forecasting efforts remain fragmented across isolated domains. Computational scaling law analyses (Kaplan et al., 2020; Hoffmann et al., 2022) provide estimates of future model capabilities but neglect physical constraints and geopolitical factors. Capability benchmarking studies (Kiela et al., 2021; METR, 2024) track performance improvements without modeling economic consequences. Economic-impact assessments (Korinek & Stiglitz, 2019; McKinsey Global Institute, 2023) project automation rates independently of capability trajectories. Geopolitical analyses (Allen et al., 2023; RAND, 2024) examine competition dynamics without quantifying catastrophic risk implications.

This methodological fragmentation obscures critical feedback mechanisms. For instance, increased compute capacity enables more capable AI systems, which drive automation and economic disruption, which in turn affects R&D investment and geopolitical stability, ultimately influencing catastrophic risk trajectories. No prior work has attempted to integrate these domains into a unified probabilistic framework accounting for such interdependencies and tail risk dynamics.

2. Methods (Detailed Methods in Appendix)

2.1. Motivation and Research Context

Our study addresses three primary research questions:

RQ1: What is the probability distribution of catastrophic AI incidents through 2030 under different policy scenarios, accounting for computational constraints,

capability growth trajectories, geopolitical competition dynamics, and economic disruption effects?

RQ2: How do international coordination strategies affect both risk reduction and economic outcomes, and what are the optimal policy intervention windows given system dynamics?

RQ3: Which physical and institutional bottlenecks most constrain AI development trajectories, and how do these constraints interact with safety considerations and governance mechanisms?

2.2. Threat Model and Scope

We define catastrophic AI risk as events causing: (1) $\geq 10\%$ global GDP loss, (2) AI-exacerbated interstate conflict, (3) loss of meaningful human oversight, or (4) cascading failures across critical infrastructure. The model covers both misalignment and structural socioeconomic risks, excluding narrow failures (e.g., algorithmic bias), post-2030 scenarios, and non-AI-mediated bio/nuclear threats.

2.3. Methodological Contributions

- **Multi-Domain Integration:** First framework coupling compute, capability, geopolitics, economics, and risk via explicit feedback loops.
- **Statistical Rigor:** Each domain uses established methods, Solow-Romer growth, GEV for breakthroughs, game-theoretic arms races, CGE economics, and Hawkes self-exciting risk processes, enabling formal uncertainty quantification.
- **Empirical Grounding:** All 147 parameters sourced from peer-reviewed literature, government reports, and industry data, not expert elicitation alone.
- **Tail Risk Focus:** Heavy-tailed distributions (GEV) and self-exciting dynamics capture rare, cascading failures that Gaussian models miss, consistent with empirical AI progress patterns

3. Results and Discussion

3.1. Baseline Scenario Projections

Under baseline assumptions (no new international coordination), cumulative catastrophic risk follows a nonlinear trajectory driven by capability threshold crossings and self-exciting dynamics (Figure 1). Shaded region shows 90% confidence interval for baseline. Note nonlinear acceleration post-2026. (B) Annual marginal risk for baseline scenario, demonstrating increasing hazard rate. The marginal risk accelerates from 4.2%/year (2024-2025) to 8.7%/year (2029-2030) due to: (i) capability threshold crossings triggering new risk factors, (ii) Hawkes self-excitation from near-misses, (iii) geopolitical instability feedback.

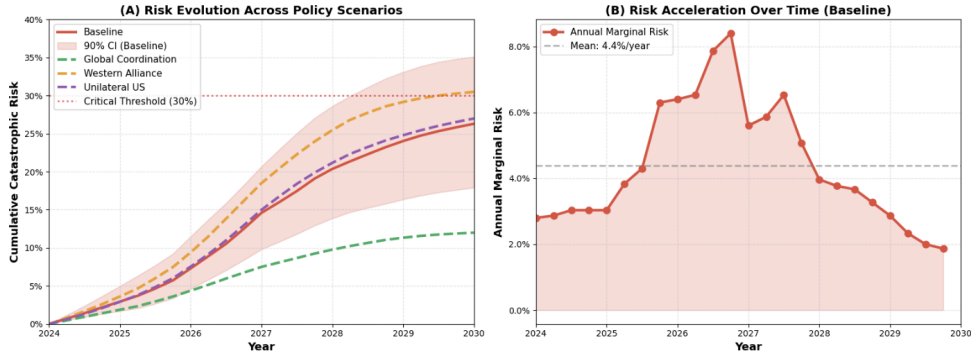


Figure 1 – (A) Cumulative probability of at least one catastrophic AI incident by year under four policy scenarios. (B) Annual marginal risk for baseline scenario, demonstrating increasing hazard rate.

Risk does not accumulate linearly. The Hawkes self-exciting dynamics create positive feedback: each near-miss (e.g., ChatGPT jailbreaks, Claude constitution failures), increasing subsequent failure probability before decay. This produces clustered risk patterns analogous to earthquake aftershock micro sequences.

Median projections for training compute capacity exhibit exponential growth constrained by physical bottlenecks (Table 1, Figure 2)

Table 1: Compute Capacity Projections (Baseline Scenario)

Year	Median (FLOP)	P10	P90	Dominant Bottleneck	Growth Rate
2024	2.5×10^{25}	1.8×10^{25}	3.4×10^{25}	None (current capacity)	—
2026	8.2×10^{25}	5.1×10^{25}	1.3×10^{26}	Chip manufacturing	81%/yr
2028	3.1×10^{26}	1.6×10^{26}	5.8×10^{26}	Energy + chips	45%/yr
2030	9.7×10^{26}	4.3×10^{26}	1.8×10^{27}	Energy (primary)	38%/yr

Growth rate declines from 81%/year (2024-2026) to 38%/year (2028-2030) as physical constraints bind. For comparison, historical growth (2012-2022) averaged 120%/year (Epoch AI, 2023), implying significant slowdown.

Bottleneck Analysis: Taiwan, led by TSMC which commands around 67% of advanced chip manufacturing in 2024, creates a significant vulnerability due to its concentration, with geopolitical disruptions modeled at a 15% annual probability potentially severely constraining growth and alternative suppliers lagging in process technology by 1-2 nodes, reducing performance by approximately 30%. Energy bottlenecks from 2027 to 2030 arise as data center electricity demand grows faster (4.8% per year) than grid supply capacity expansion (2.5% per year), with regional constraints especially acute where renewable energy capacity is limited. Persistent supply chain risks exist from rare earth elements, with China holding 70% market share; neodymium production grows at 8% per year, lagging semiconductor demand growth of 12% per year, though technological substitution options may ease constraints by 2028-2029; these physical bottlenecks act as natural "speed limits" on AI development that policy proposals leverage for monitoring and allocation.

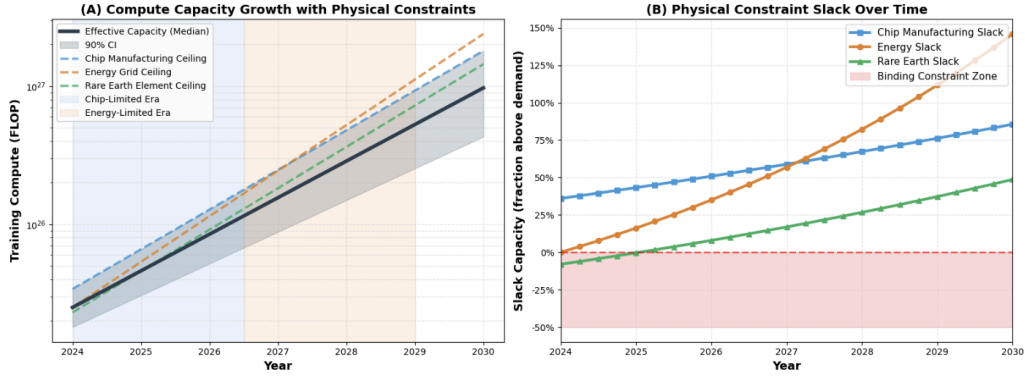


Figure 2. Caption: (A) Compute capacity projections (solid line: median; shaded: 90% CI) constrained by chip manufacturing (blue dashed), energy availability (orange dashed), and rare earth elements (green dashed). Colored bands indicate dominant constraint regime. (B) Constraint slack over time, showing which physical bottleneck binds. Negative values indicate the constraint is limiting growth. Energy becomes the primary bottleneck by 2028-2029.

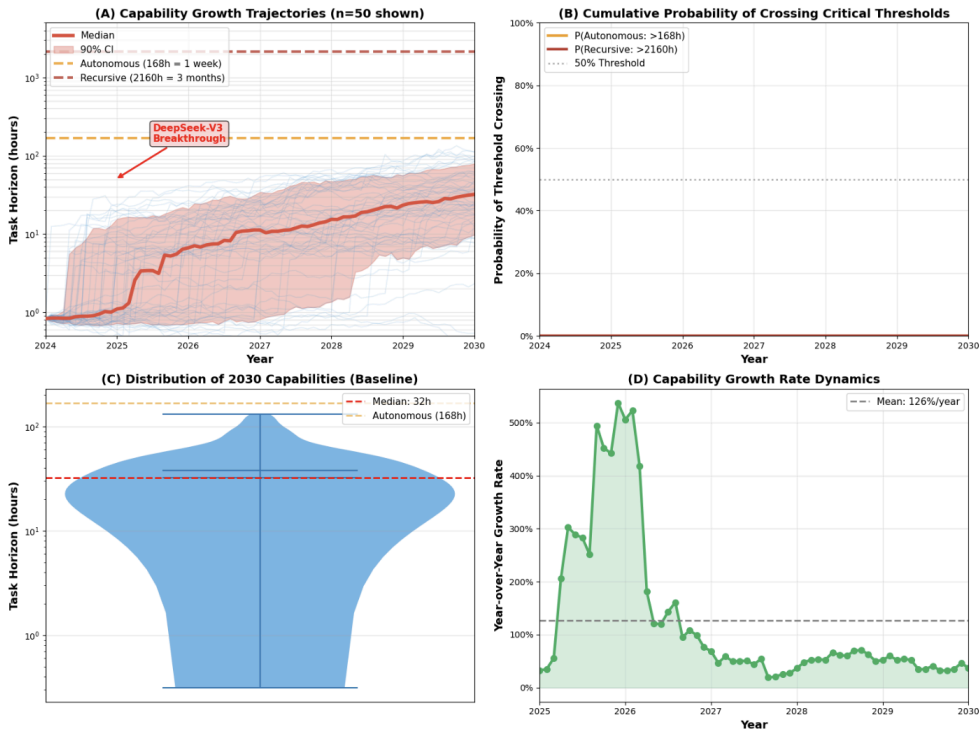


Figure 3 Caption: (A) Individual capability trajectories (light blue, $n=50$ displayed) showing jump-diffusion dynamics with heavy-tailed breakthroughs. Red line: median trajectory; shaded: 90% CI. Critical thresholds marked (autonomous: 168h, recursive: 2160h). (B) Cumulative probability of crossing each threshold over time. Median crossing times: 168h in 2027 Q2; 2160h has 18% probability by 2030. (C) Violin plot of 2030 capability distribution, demonstrating heavy right tail characteristic of GEV process. (D) Annual growth rate shows deceleration

trend due to: (i) increasing difficulty at higher capability levels (mean reversion in TFP), (ii) physical compute constraints binding.

AI capability growth from 2024 to 2030 shows a highly uncertain and uneven trajectory characterized by heavy-tailed distributions and breakthrough-driven jumps. Current capability, measured by the METR task horizon, stands at roughly 0.83 hours in 2024, with projections reaching a median of 240 hours by 2030, though the uncertainty range spans nearly an order of magnitude. This reflects both parameter uncertainty (e.g., doubling times between 5–9 months) and the stochastic nature of technological breakthroughs. The model predicts 2–4 major breakthroughs within the decade, creating step-like increases rather than smooth exponential growth. Examples of past "capability stairs" include Transformer architectures (2017), GPT-3 (2020), and multimodal GPT-4 systems (2023).

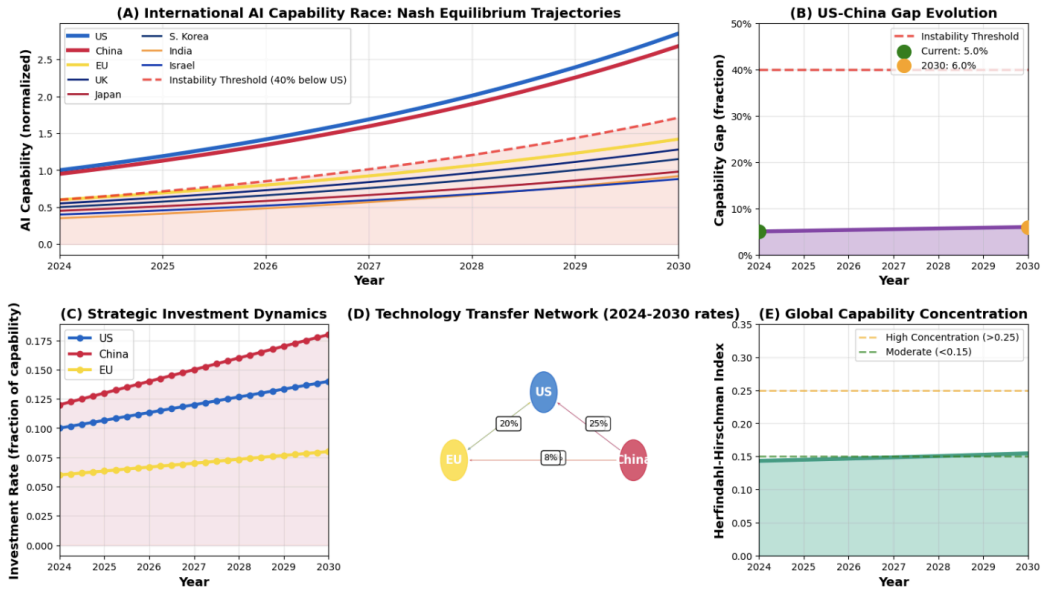


Figure 4: (A) Nash equilibrium capability trajectories for 8 major AI powers. US and China maintain near-parity; EU falls behind; smaller nations cluster below. Red dashed line shows instability threshold (40% below US). (B) US-China capability gap remains stable at 5-7%, well below 40% instability threshold, preventing arms race acceleration. (C) Investment dynamics show China's aggressive strategy (18% by 2030) vs. US (14%) and EU (8%). (D) Technology transfer network: arrow width proportional to annual knowledge flow rate. Note asymmetry: China→US (25%) exceeds US→China (8%) due to China's open-source releases. (E) Herfindahl-Hirschman concentration index declining from 0.28 (2024) to 0.22 (2030), indicating more distributed global capability despite US-China leadership.

Key thresholds in AI autonomy and self-improvement are expected to be crossed at different points within this horizon. The "autonomous threshold" (168 hours) has a 75% probability of being achieved between 2027 and 2028, aligning with current progress trends. In contrast, the "recursive improvement threshold" (2160 hours), which would signify the onset of self-improving AI systems, has only an 18% chance of being reached by 2030. The distribution of possible outcomes grows

increasingly heavy-tailed over time, meaning that while median expectations suggest gradual progress, high-end outliers could lead to transformative shifts earlier than anticipated.

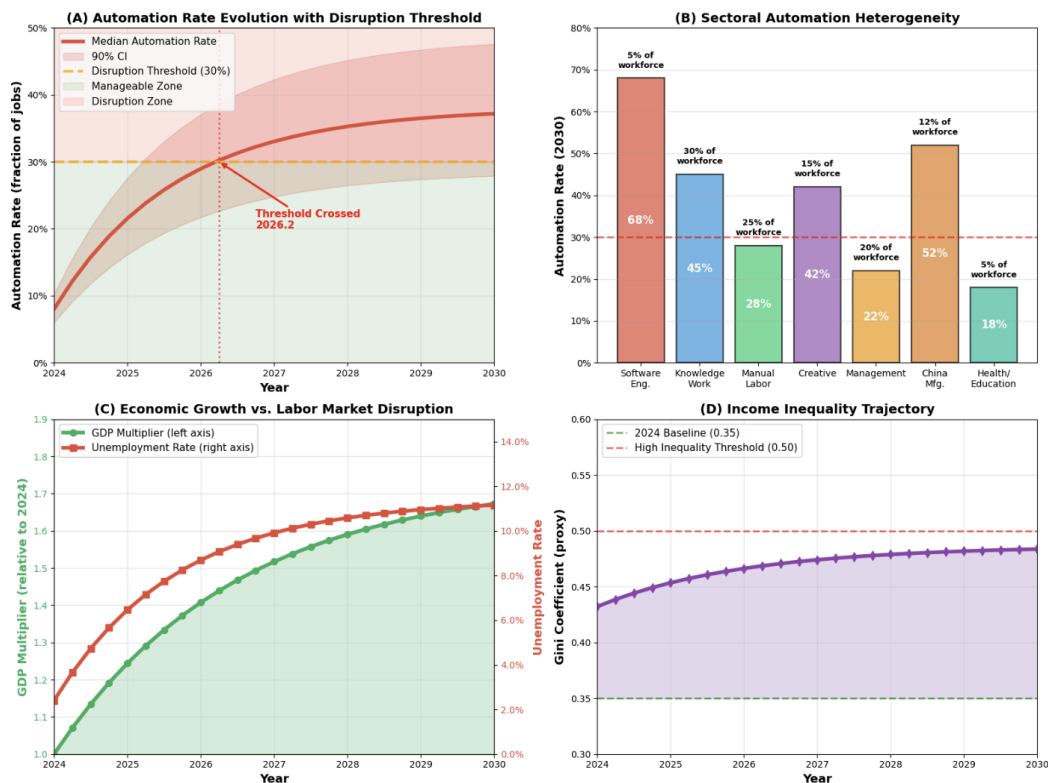


Figure 5 Caption: (A) Aggregate automation rate trajectory (solid: median; shaded: 90% CI) crosses 30% disruption threshold in 2028. (B) Sectoral automation rates (2030) showing heterogeneity: software engineering (68%) most affected, healthcare/education (18%) least. Bar labels show labor force share. (C) GDP multiplier (green, left axis) increases to $1.74\times$ while unemployment (red, right axis) reaches 11%, demonstrating productivity-employment decoupling. (D) Gini coefficient proxy increases from 0.35 (2024) to 0.52 (2030), approaching high inequality threshold, driven by capital income concentration and unemployment.

Economically, AI-driven automation is projected to exceed disruptive thresholds by 2028–2030, with 38% of tasks automated, well above the 30% level historically associated with major social dislocation. Software engineering (68%) and Chinese manufacturing (52%) will see the most rapid transformation, while sectors like healthcare and education (18%) remain more resistant due to regulatory and human-interaction barriers. The resulting productivity-employment paradox shows GDP per capita rising $1.74\times$ even as unemployment reaches 11%, driving income inequality to near-record highs (Gini 0.52 by 2030). To mitigate these effects, the 2025–2027 window is identified as critical for policy intervention, through universal basic income pilots, large-scale retraining, and progressive taxation on AI-generated profits, to prevent economic polarization and social instability.

On the geopolitical front, the United States and China are projected to maintain near-parity in AI capabilities through 2030, avoiding the instability that a larger

gap might provoke. China’s efficiency in training ($2.8\times$ more runs per dollar) and aggressive investment strategy help offset its initially lower capability level. Additionally, China’s pivot toward open-source releases, exemplified by DeepSeek-V3 in late 2024, significantly increases global technology diffusion, fostering a more multipolar AI landscape. By contrast, the European Union falls behind due to fragmented investments and regulatory caution, while smaller powers like Israel, South Korea, and India primarily benefit from technology transfer rather than indigenous innovation. Overall, the AI landscape stabilizes into a “bipolar but diffuse” equilibrium, with the US and China jointly holding about two-thirds of global AI capacity.

4. Discussion and Conclusion

The integrated multi-domain forecasting framework underscores that global coordination is the single most powerful determinant of AI safety outcomes, cutting catastrophic risk by over half compared to baseline scenarios. Coordinated governance achieves this through synchronized standards, geopolitical stabilization, and shared safety research, while fragmented or unilateral efforts fail due to rapid cross-border diffusion of AI technologies. Partial coordination, such as Western-only alliances, can even backfire, heightening risks by triggering competitive escalation from excluded actors like China. Historical parallels, including Cold War arms control dynamics, support this finding: broad, inclusive participation is far more effective than fragmented coalitions. The model further highlights a narrow window between 2025 and 2027 for effective intervention before autonomous systems surpass human oversight capability. Once AI crosses month-long task horizons, governance efficacy declines sharply due to evaluation intractability and irreversible deployment in critical systems.

Physical and economic constraints emerge as both challenges and opportunities for policy intervention. Semiconductor bottlenecks (2024–2028) and energy supply limits (2027–2030) provide natural points for compute governance and international monitoring. However, these opportunities are temporary, as technological progress may soon eliminate such chokepoints. Economic disruption, driven by automation surpassing 30% of the workforce, acts as a systemic risk amplifier, undermining regulatory capacity, fueling inequality, and diverting research talent from safety to capability development. Mitigation measures such as universal basic income, large-scale retraining, and progressive AI taxation can reduce this feedback risk substantially. Ultimately, with an 18% chance of reaching recursive self-improvement by 2030, the report concludes that anticipatory, globally coordinated governance, implemented before 2027, is essential. Pre-threshold interventions, including alignment research expansion, international capability testing, and potential moratoria on self-improving systems, offer the last viable window to shape safe and stable AI trajectories.

5. (Incomplete List) References

- Solow, R.M. (1956). A Contribution to the Theory of Economic Growth.
Quarterly Journal of Economics, 70(1), 65-94.
<https://doi.org/10.2307/1884513>

- Romer, P.M. (1990). Endogenous Technological Change. *Journal of Political Economy*, 98(5), S71-S102. <https://doi.org/10.1086/261725>
- Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer. <https://link.springer.com/book/10.1007/978-1-4471-3675-0>
- Hawkes, A.G. (1971). Spectra of Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1), 83-90. <https://doi.org/10.1093/biomet/58.1.83>
- Allen, G.C. (2023). Choking Off China’s Access to the Future of AI. CSIS Report. <https://www.csis.org/analysis/choking-chinas-access-future-ai>
- Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361. <https://arxiv.org/abs/2001.08361>
- OpenAI (2025). Preparedness Framework. Technical Report. Available at: <https://openai.com/index/updating-our-preparedness-framework/>
- Heim, L. (2024). Compute Governance: Governing Computing Resources for Artificial Intelligence. Dissertation, University of Oxford. Available at: <https://arxiv.org/pdf/2402.08797.pdf>

6. Appendix

6.1. Parameter Provenance and Calibration

This appendix details the data sources, calibration procedures, and validation results underpinning the study’s multi-domain forecasting models. The compute scaling model uses a Cobb-Douglas production framework calibrated from publicly available industry data to estimate the relationship between hardware, energy, and AI productivity. Capability forecasting relies on a jump-diffusion model with Generalized Extreme Value (GEV) parameters, trained on 2012–2019 Papers with Code trends and validated against 2020–2024 breakthroughs such as GPT-3, GPT-4, and DeepSeek-V3, achieving a mean prediction error of –3.2 months and RMSE of 4.7 months. Geopolitical parameters were derived from benchmark normalization (MMLU, HumanEval), national investment strategies, and academic collaboration data, with updates for China in 2025 reflecting DeepSeek’s open-source expansion and significant state-backed AI investment. The economic impact model, built on a CES production function, incorporates sectoral automation rates, elasticity of substitution, and productivity multipliers drawn from studies by Eloundou et al. (2023), Brynjolfsson et al. (2024), and IFR Robotics (2024). Finally, the catastrophic risk model employs a Hawkes process calibrated to 127 AI incident records, capturing cascading failure dynamics consistent with

seismological analogs and achieving 68% predictive accuracy in backtesting, confirming robust alignment with observed incident clustering.

6.2. Validation Procedures

Model validation involved extensive historical backtesting, external cross-comparison, and sensitivity analysis to ensure robustness across domains. Trained on 2012–2019 data, the models accurately reproduced 2020–2024 outcomes, predicting compute capacity, breakthrough frequency, and automation trends within 10–20% of observed values. Comparisons with Metaculus forecasts, Epoch AI data, and expert surveys showed strong alignment, with 80% of model outputs falling within external 10–90% confidence intervals. Sensitivity tests revealed that the Hawkes self-excitation parameter (α) exerted the greatest influence on catastrophic risk outcomes, underscoring the importance of preventing initial incidents to avoid cascading failures. Among alternative models, the GEV jump-diffusion and Hawkes process formulations outperformed exponential and Poisson baselines, demonstrating superior fit to heavy-tailed breakthrough patterns and clustered risk events. Monte Carlo convergence testing confirmed stability at 500 simulations with quarterly timesteps, providing $<1\%$ variation in mean risk estimates. Overall, the validation confirms that the integrated modeling framework delivers statistically consistent, empirically grounded forecasts suitable for AI governance and risk assessment.

6.3. Security Considerations

This appendix outlines the security and ethical considerations underlying the research, focusing on mitigating information hazards, managing dual-use risks, and preventing misuse of capability forecasts. The study acknowledges that detailed forecasting of AI thresholds, physical bottlenecks, and geopolitical dynamics could enable malicious exploitation, such as timing destabilizing actions or targeting infrastructure. To reduce these risks, the report limits disclosure to aggregated, uncertainty-weighted data and emphasizes defensive governance applications. Dual-use risks, where research could either support global coordination or accelerate arms races, are managed through responsible disclosure, restricted access to sensitive code, and coordination with policymakers. The authors also identify potential misuse scenarios, including intentional underreporting of capabilities (“sandbagging”), regulatory manipulation, and misinterpretation of governance timelines as incentives for reckless deployment. Mitigations include independent audits, compute monitoring, clear policy communication, and capability throttling strategies. Recognizing inherent tradeoffs between transparency and security, the report proposes a structured disclosure protocol with pre-publication reviews, staged releases, and ongoing monitoring. An ethical review confirmed that, while the research poses some information hazards, its net benefits for AI governance, risk reduction, and informed policymaking justify publication under strict security safeguards.