Stop Predicting AI. Start Preparing for It.

Deriving government preparedness from AI risk scenarios

The future of advanced AI is highly uncertain, spanning a broad range of possible paths, some of which pose severe risks. Across jurisdictions, government preparations include:

- forming AI safety institutes that evaluate models for dangerous capabilities[1;2];
- requiring or encouraging frontier safety frameworks that link safeguards to capability thresholds[6;8;9;14;17];
- ranking AI risks by likelihood and impact in national risk assessments and registers[3;13];
- drafting response plans for specific AI threat models[5;19;21].

Most of these efforts follow a "predict-then-act" paradigm[16], but AI defies prediction: experts' timelines differ by orders of magnitude[11], capabilities emerge abruptly[22], and failure modes surprise even their developers[10].

When predictions fail, capabilities built narrowly around them fail, too[7]. An influenza pandemic topped the UK risk register for a decade, but preparations missed the asymptomatic transmission that defined COVID-19[18;20]. Advanced AI's risks are harder still to predict. They require a different kind of preparation.

Governments have faced deep uncertainty before: defence[7], flood[12] and financial supervisory policy[4] no longer prepare for a *single predicted future* but build capabilities that offer *protection across many*. This works because defining capabilities remains tractable even after accurate risk assessment becomes impossible[15;16].

This approach, known as "monitor-and-adapt", requires mapping a representative range of risk scenarios to the government capabilities they demand—to reveal which provide value across many futures or cost-effective insurance in the most severe cases[12;16] (Fig. 1).

No such mapping exists yet for AI[21]. We are building it as a public resource in three layers: a **library** of AI risk scenarios, combining literature with systematic scenario generation; an **assessment** of which capabilities each scenario demands, from technical containment to institutional coordination and market regulation, with a focus on where government capacity is least established; and a **prioritisation** to identify which capabilities governments most need.

HELP US BUILD IT

The **scenario library and capability taxonomy** are open for comment. We invite you to:

- **Join the assessment.** Which capabilities does each scenario demand?
- **Challenge the scenario set.** What is our library missing?
- **Critique the method.** Where is more work needed and what are limitations?

Get in touch: isaak.mengesha@arcadiaimpact.org

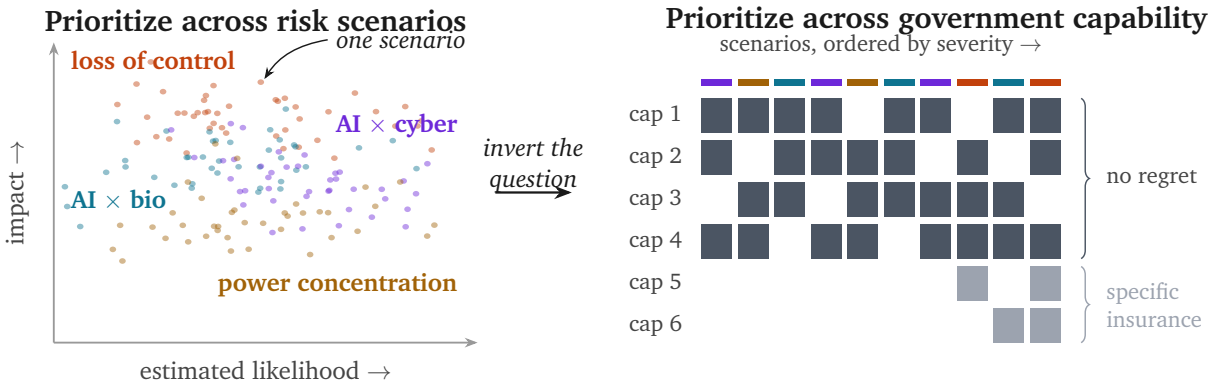

Figure 1: Inverting the question. *Left:* AI risk scenarios ranked by likelihood and impact—the predict-then-act input, where estimates diverge by orders of magnitude. *Right:* the same scenarios as columns, government capabilities as rows: those demanded across many scenarios form a "no-regret" core; those demanded by few, but severe, scenarios are insurance.

References

[1] Gregory C. Allen and Georgia Adamson. The AI safety institute international network: Next steps and recommendations. Report, Center for Strategic and International Studies (CSIS), Washington, D.C., October 2024. URL https://csis-website-prod.s3.amazonaws.com/s3fs-public/2024-10/241030_Allen_Safety_Network.pdf. Accessed 2026-07-20.

[2] Peter Barnett and Aaron Scher. AI governance to avoid extinction: The strategic landscape and actionable research questions. *arXiv preprint arXiv:2505.04592*, 2025.

[3] David Blagden. The flawed promise of national security risk assessment: Nine lessons from the British approach. *Intelligence and National Security*, 33(5):716–736, 2018. doi: 10.1080/02684527.2018.1449366.

[4] Board of Governors of the Federal Reserve System. The supervisory capital assessment program: Overview of results. Technical report, Board of Governors of the Federal Reserve System, Washington, D.C., May 2009. URL https://www.federalreserve.gov/newsevents/files/bcreg20090507a1.pdf. Published 7 May 2009. First post-2008 U.S. supervisory stress test of the 19 largest bank holding companies.

[5] Benjamin Boudreaux, Michael J. D. Vermeer, Kamaria Horton, and Nidhi Kalra. The case for AI loss of control response planning and an outline to get started. Expert Insights PE-A4232-1, RAND Corporation, Santa Monica, CA, October 2025. URL https://www.rand.org/pubs/perspectives/PEA4232-1.html.

[6] California State Legislature. SB-53: Artificial intelligence models: Large developers (transparency in frontier artificial intelligence act). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53, September 2025.

[7] Paul K. Davis. Analytic architecture for capabilities-based planning, mission-system analysis, and transformation. Technical Report MR-1513-OSD, RAND Corporation, Santa Monica, CA, 2002. URL https://www.rand.org/pubs/monograph_reports/MR1513.html.

[8] European Commission. General-purpose AI code of practice. European AI Office. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai, July 2025.

[9] Frontier Model Forum. Introducing the FMF's technical report series on frontier AI frameworks. https://www.frontiermodelforum.org/updates/introducing-the-fmfs-technical-report-series-on-frontier-ai-safety-frameworks/, April 2025. Overview of frontier AI safety frameworks; notes 12 developers have published frameworks following the May 2024 Seoul Frontier AI Safety Commitments.

[10] Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, Dario Amodei, and Jack Clark. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 2022. doi: 10.1145/3531146.3533229. URL https://arxiv.org/abs/2202.07785.

[11] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. Thousands of AI authors on the future of AI. https://arxiv.org/abs/2401.02843, 2024. 2,778 researchers; timeline estimates span decades with large inter-expert variance.

[12] Marjolijn Haasnoot, Jan H. Kwakkel, Warren E. Walker, and Judith ter Maat. Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world. *Global Environmental Change*, 23(2):485–498, 2013. doi: 10.1016/j.gloenvcha.2012.12.006.

[13] HM Government, Cabinet Office. National risk register 2025. HM Government. https://www.gov.uk/government/publications/national-risk-register-2025, January 2025. Published 16 January 2025. Reasonable-worst-case scenarios across 89 acute risks; likelihood-impact matrix; dynamic assessment process introduced in the 2025 edition.

[14] Leonie Koessler, Jonas Schuett, and Markus Anderljung. Risk thresholds for frontier AI. *arXiv preprint arXiv:2406.14713*, 2024. doi: 10.48550/arXiv.2406.14713. URL https://arxiv.org/abs/2406.14713.

[15] Robert J. Lempert, Steven W. Popper, and Steven C. Bankes. Shaping the next one hundred years: New methods for quantitative, long-term policy analysis. Technical Report MR-1626-RPC, RAND Corporation, Santa Monica, CA, 2003. URL https://www.rand.org/pubs/monograph_reports/MR1626.html.

[16] Vincent A. W. J. Marchau, Warren E. Walker, Pieter J. T. M. Bloemen, and Steven W. Popper, editors. *Decision Making under Deep Uncertainty: From Theory to Practice*. Springer, Cham, Switzerland, 2019. doi: 10.1007/978-3-030-05252-2. URL https://link.springer.com/book/10.1007/978-3-030-05252-2. Open access. Chapter 1 (Marchau, Walker, Bloemen & Popper) introduces the predict-then-act vs. monitor-and-adapt paradigms and the DMDU decision-support framework.

[17] METR. Common elements of frontier AI safety policies. Technical report, Model Evaluation and Threat Research (METR), March 2025. URL https://metr.org/common-elements. Documents shared use of capability thresholds, evaluations, and pre-committed mitigations across Anthropic, OpenAI, and Google DeepMind frameworks.

[18] John Raine. Half of the national risk register is missing. *RUSI Newsbrief*, 41(1), January 2021. URL https://www.rusi.org/explore-our-research/publications/rusi-newsbrief/half-national-risk-register-missing. Royal United Services Institute. Argues that ranking threats by statistical likelihood is mistaken for managing them.

[19] The White House. Winning the race: America's AI action plan. Executive Office of the President. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf, July 2025. Published 23 July 2025. Under "Promote Mature Federal Capacity for AI Incident Response," recommends modifying existing incident-response frameworks (e.g., CISA Cyber Incident & Vulnerability Response Playbooks) to incorporate AI. A blueprint of recommendations, not binding mandates.

[20] UK Covid-19 Inquiry. Module 1: The resilience and preparedness of the united kingdom. Technical report, UK Covid-19 Inquiry, London, July 2024. URL https://covid19.public-inquiry.uk/reports/module-1-report-the-resilience-and-preparedness-of-the-united-kingdom/. Chair: The Rt Hon the Baroness Hallett DBE. Published 18 July 2024. First report; found the UK was under-prepared for the pandemic.

[21] Akash Wasil, Everett Smith, Corin Katzke, and Justin Bullock. AI emergency preparedness: Examining the federal government's ability to detect and respond to AI-related national security threats. https://arxiv.org/abs/2407.17347, 2024. Three scenarios (loss of control, weight theft, bio); evaluates federal detect/prevent/respond gaps.

[22] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2206.07682.