How to prepare for AI crises you cannot predict

AT A GLANCE

- Most government AI preparation follows a predict-then-act logic: rank threats by likelihood, prepare for the top of the list. AI resists prediction, and capabilities built around a failed forecast fail with it.
- We adapt capability-based planning from defence and homeland security: sample a scenario library across declared axes, rate government capabilities against every scenario, and extract priorities under explicit decision rules.
- Piloted across the four most severe AI threat classes: robust priorities surface even from imperfect ratings, before likelihood or risk preference is settled.

Across jurisdictions, governments are preparing for advanced AI: safety institutes evaluate frontier models for dangerous capabilities[1;3], safety frameworks link safeguards to capability thresholds[10;15;18], national risk registers rank AI risks by likelihood and impact[14], serious-incident reporting is entering into force (EU AI Act Article 73 since August 2026, California's SB-53 since January)[7;9], and response plans are being drafted for specific threat models[6;20]. Every one of these assumes that emergent AI threats can be identified, accurately ranked, and met with a suitable strategy before they cause harm.

Other fields facing deep uncertainty gave that assumption up. United States defence planning stopped preparing against a single predicted adversary two decades ago[8]; financial supervisors stress-test banks against adverse scenarios rather than forecasts[5]; Dutch flood policy builds adaptive pathways instead of betting on one water level[13]. Each asks what capabilities must hold across many futures, not which future is likeliest. AI governance has not yet made this move.

The predictions it rests on fail routinely. An influenza pandemic topped the United Kingdom's risk register from 2008 to 2020, yet preparations missed the asymptomatic transmission that defined COVID-19, and the response failed[19;21]. AI is harder still: expert timelines disagree by orders of magnitude[12], capabilities emerge abruptly[23], and failure modes surprise the developers themselves[11]. The AI preparedness work that does exist goes deep on single threats, a loss-of-control response plan here[6], a federal gap assessment there[22], and stays silent on the rest.

The deeper problem is that likelihood and preparation are different questions which current planning treats as one. Even the most probable scenario is individually unlikely to arrive as predicted, or at all, so narrow scenario-specific investment is a poor bet[16]. Meanwhile the governance instruments now arriving presuppose a capable counterpart: developer if-then commitments need someone to verify thresholds and enforce them[2;15], and incident-reporting duties need a recipient that can attribute, contain, and remediate what is reported[7;9]. What the state must be able to do is specified nowhere. That question, unlike likelihood, stays tractable under deep uncertainty.

Fortunately, this is a familiar problem in policy analysis, formalised in the field of Decision Making under Deep Uncertainty[17]. Applying its principles to AI, we developed a methodological framework in three parts:

1. **Definition:** a "scenario library" samples systematically across precommitted binary

axes (intent, legitimacy, autonomy, concentration) rather than by expected harm; three axes per threat class give eight scenarios, so selection bias is declared rather than hidden.

2. **Assessment:** a "capability", a pair of function and target (contain × compute, verify × models), is rated against every scenario through coarse, gated criteria: applicable, deployable ad hoc, necessary, then effectiveness and externalities, with reasons recorded before scores.
3. **Prioritisation:** the resulting matrix is read through whichever decision rule a government holds: no-regret, decisive-only, do-no-harm, load-bearing, or worst-case insurance. Capabilities demanded across many futures form a core; those demanded only by the severest scenarios are cheap insurance.

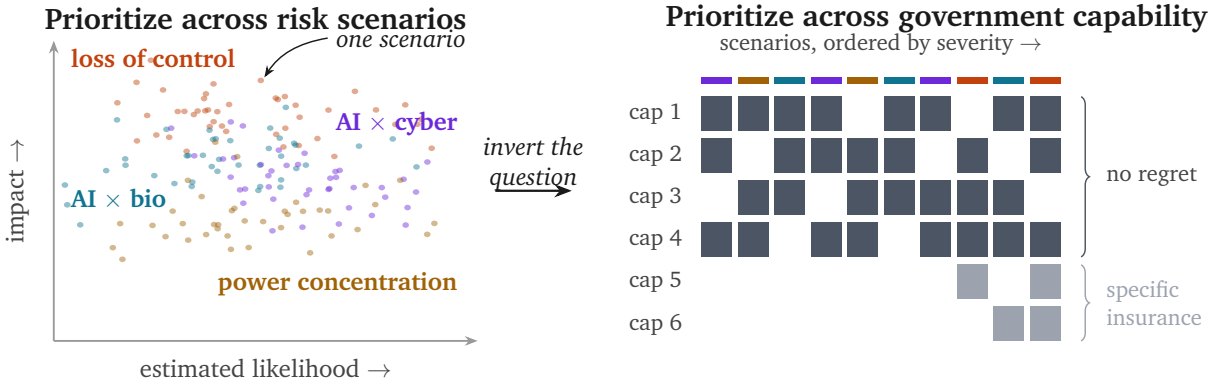

Figure 1: Inverting the question. *Left:* AI risk scenarios ranked by likelihood and impact, the predict-then-act input, where estimates diverge by orders of magnitude. *Right:* the same scenarios as columns, government capabilities as rows: those demanded across many scenarios form a "no-regret" core;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

We piloted the framework on the four threat classes the literature ranks most severe: loss of control, power concentration, and AI-enabled biological and cyber attack, 29 scenarios in all, with 13 of 65 capabilities rated so far, each at most twice. The numbers are illustrative, but three patterns recur. First, almost no rated capability can be fielded mid-crisis from generic government capacity: the pre-positioning gap runs from 36% of demanded cases in cyber to 62% in loss of control. Second, prioritisation is largely stable across decision rules: institutional mobilisation and international joint action lead nearly every column, meaning a government could act on them before settling its risk preference, defusing the threshold disagreements that block coordinated action elsewhere[4]. Third, tail insurance is short and concrete: the capabilities demanded only by existential scenarios reduce to one containment family spanning compute, models, deployment, and infrastructure.

The instrument maps demand, not capacity: "deployable" assumes a generically competent government, while actual endowments vary by jurisdiction and need their own audit. Ratings encode rater priors, the library is a design choice a different team would draw differently, and the coarse gates yield an ordering rather than a resource allocation. It predicts nothing and triggers nothing; it says what to hold, not when to invest.

A naive reading of the AI risk debate would start preparedness where public concern is loudest, with biological and cyber misuse. The matrix reads differently. Bio and cyber scenarios largely tap existing institutional machinery: 38% and 36% of demanded

capability cannot be improvised. Loss of control and power concentration invert this: 62% and 57% must be pre-positioned, and single loss-of-control scenarios bind up to 12 of the 13 rated capabilities at once, so one missing capability voids the entire response. Preparation is needed everywhere; where it must happen before the crisis, the reading reverses.

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] Anthropic. Responsible scaling policy (version 3.0). https://www.anthropic.com/responsible-scaling-policy/rsp-v3-0, February 2026. Version 3.0, February 2026. Separates company-specific plans from industry recommendations; conditions slowdown on competitors' verifiable safety measures.

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

[4] Scott Barrett and Astrid Dannenberg. Sensitivity of collective action to uncertainty about climate tipping points. *Nature Climate Change*, 4(1):36–39, 2014.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, OJ L, 2024/1689, 12 July 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. Entered into force 1 August 2024; high-risk provisions, including Article 73 (reporting of serious incidents), apply from 2 August 2026 (Article 113).

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.