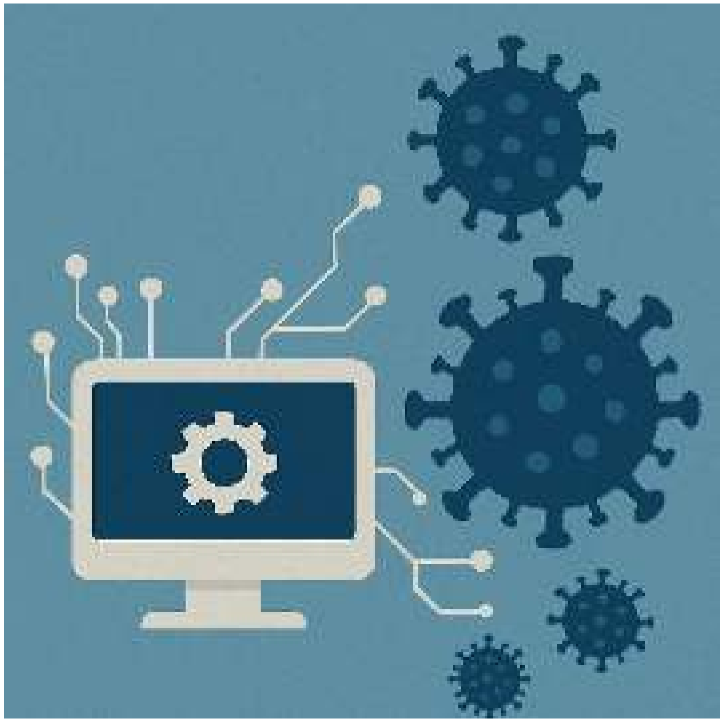
Anticipating and Managing Threats from Artificial Intelligence and Bioweapons

Prof Nick Wilson, HPARC, University of Otago Wellington
Dr Matt Boyd, Adapt Research Ltd (22 August 2025)



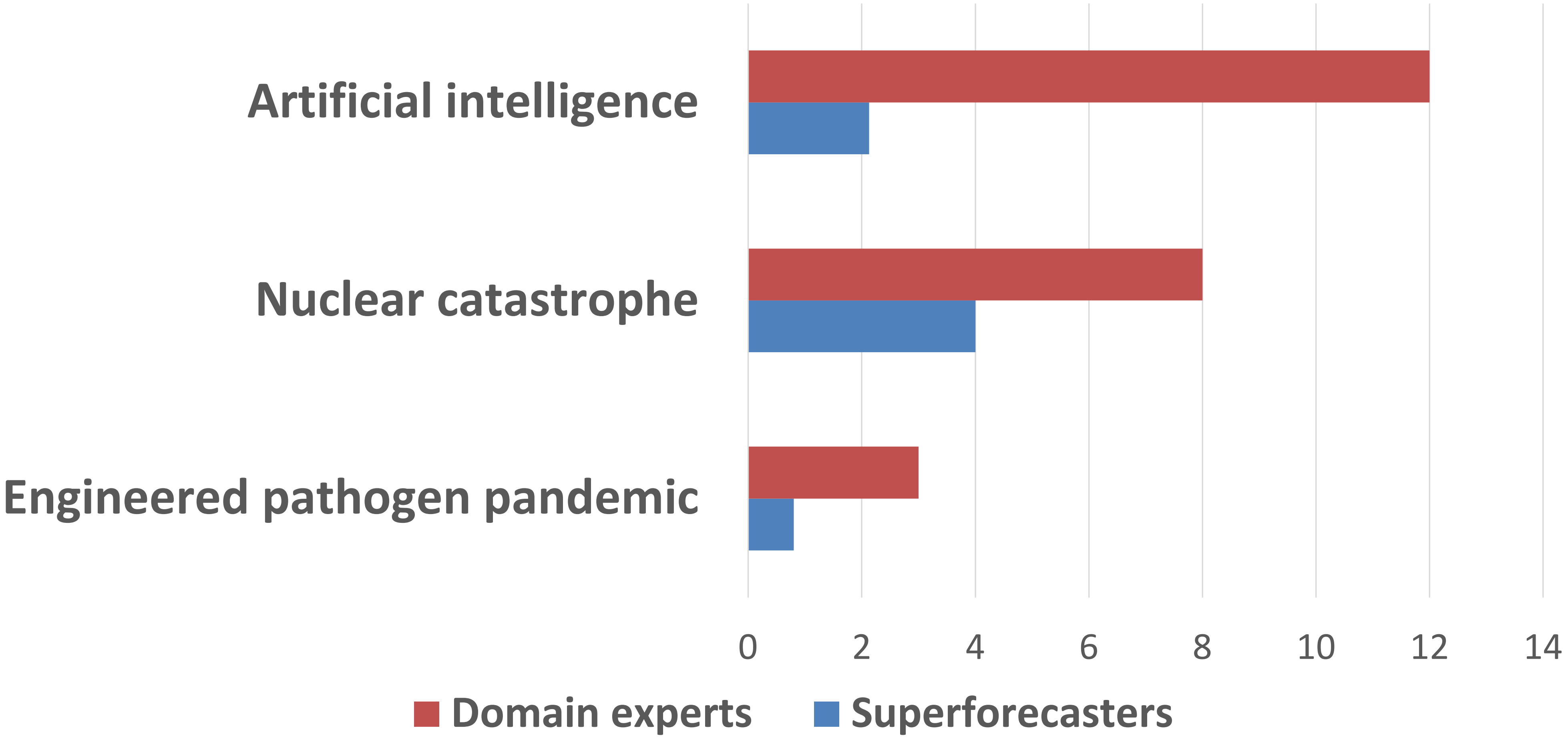
Adapt Research
As we build our world we build our minds

Many Concerns About AI

While AI appears to have enormous potential value in many fields, there are major concerns:

- Potential mass unemployment → **societal disruption**
- Risk of facilitating **state conflict** (via: robot armies, cyberattacks, nuclear weapon control systems [Nature 2025], bioengineered pandemics)
- **Societal take-over** by non-aligned AI

Risk of the Top 3 Catastrophes (% likelihood by year 2100; for 10%+ of global population killed; median values; Source: Karger et al 2023)



Since Karger et al – Risks have Likely Increased

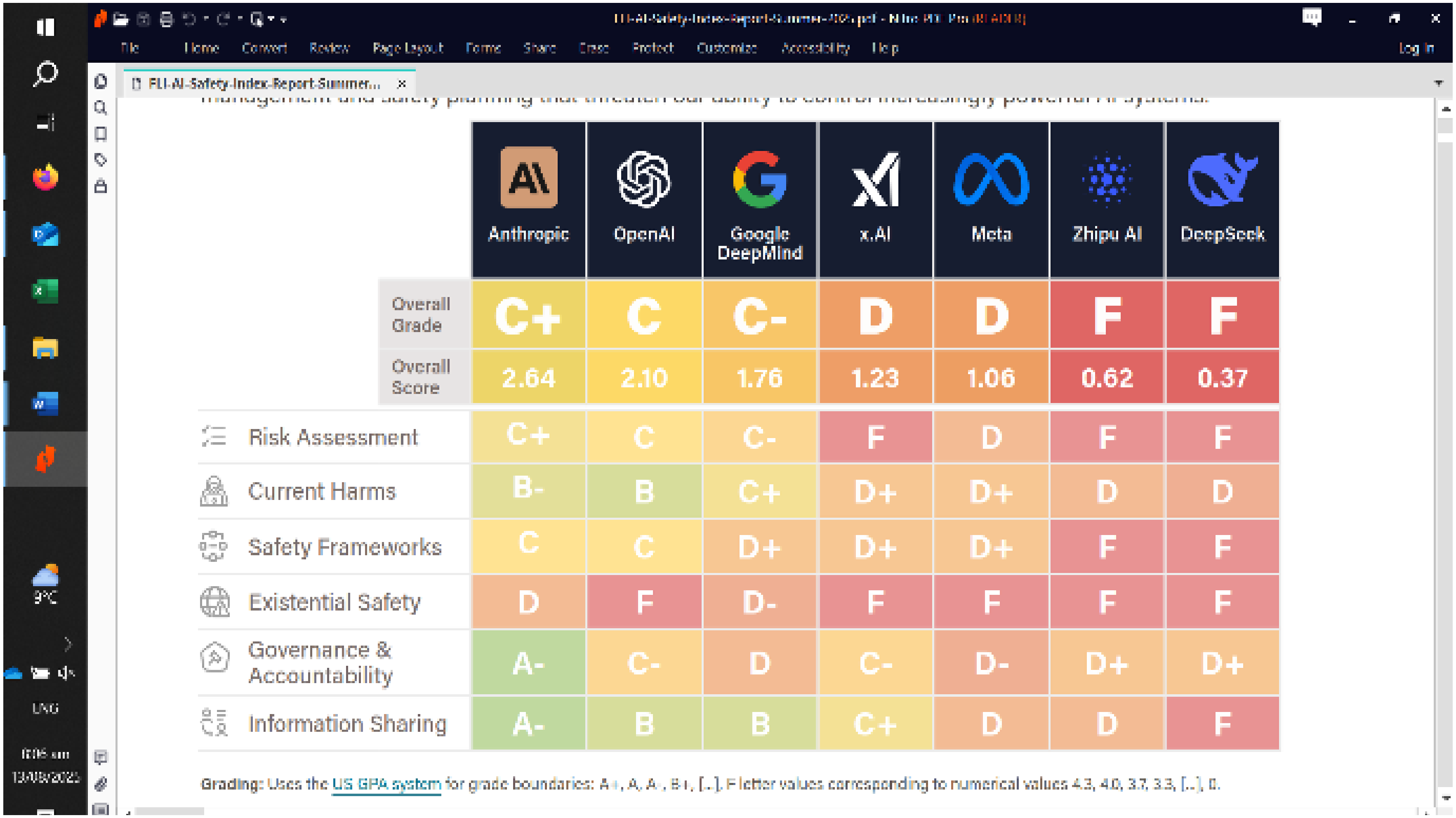
- **Progress with AI** (especially in US & China), including ongoing release of open-weight LLMs
- **Little regulation** of AI by governments or the UN (US Government in apparent AI race with China; intense competition between AI companies)
- Open **letter on existential risks** from AI (350+ experts) & several Nobel laureates; Statement by The Elders

Experts Might be Underestimating Progress

- Study by FRI (2025) with experts in biology & biosecurity (n=46) and generalist forecasters (n=22)
- Median expert predicted a human-caused epidemic (>100,000 deaths) at 1.5% conditional on several hypothetical LLM capabilities (including matching the performance of a top performing team of virologists)
- But study suggested that LLMs have already crossed this performance threshold eg, OpenAI's o3 model. Yet median respondent thought that this would not happen until after 2030

Companies score poorly on AI Safety Index

(Future of Life Institute July 2025)


FLI-AI-Safety-Index-Report-Summer-2025.pdf - Nitro PDF Pro (READER)
File
Home
Convert
Review
Page Layout
Forms
Share
Erase
Protect
Customize
Accessibility
Help
Log In
FLI-AI-Safety-Index-Report-Summer...
9°C
ENG

management and safety planning that threaten our ability to control increasingly powerful AI systems.

	Anthropic	OpenAI	Google DeepMind	x.AI	Meta	Zhipu AI	DeepSeek
Overall Grade	C+	C	C-	D	D	F	F
Overall Score	2.64	2.10	1.76	1.23	1.06	0.62	0.37
Risk Assessment	C+	C	C-	F	D	F	F
Current Harms	B-	B	C+	D+	D+	D	D
Safety Frameworks	C	C	D+	D+	D+	F	F
Existential Safety	D	F	D-	F	F	F	F
Governance & Accountability	A-	C-	D	C-	D-	D+	D+
Information Sharing	A-	B	B	C+	D	D	F

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

AI + Life Sciences = Huge Potential & Potential Catastrophic Risks

Based on: “Statement on Biosecurity Risks at the Convergence of AI and the Life Sciences” by the Nuclear Threat Initiative [NTI 2025]

- AIxBio tools simplify pathogen design and so lower barriers to designing dangerous biological agents (governments more so than non-state actors [Sandberg & Nelson 2020])
- Accidental or deliberate misuse is more feasible
- Risks of global biological catastrophes

Raising the Ceiling of Potential Harm

- AI could design pathogens **more dangerous** (virulence and/or transmission) than natural ones:
  - Produce genome sequences that encode new viruses or weaponise natural ones eg, smallpox
  - New individual biological molecules (toxins, proteins found in pathogens, or proteins that bind to important targets in the body)
  - Design of groups of biomolecules working together (eg, on cell signaling)
- **Seems plausible** within a few years without safeguards [NTI 2025]
- Release of multiple **simultaneous pandemics** in AI take-over scenario [RAND 2025]

Rise of Autonomous AI Agents

- Agent capacity being released with publicly available LLMs (but might still need new AI breakthrough for real risk?)
- Agents can assist in: literature review, hypothesis generation, design lab experiments, interface with bioscience lab equipment, & work with advanced laboratory robotics [NTI 2025]
- Could be misused or act in unintended ways
- Concurrent rapid progress in life science applications

Autonomous AI Agents – wet lab interaction

- AI can use robotics in wet lab work
- Or could manipulate human lab workers (tricking, bribes and/or blackmail). Eg, LLMs can resort to blackmail in lab experiments – when pushed [Anthropic Report 2025]

AIxBIO Undermining Biosecurity Defences

AI could contribute to:

- Circumventing nucleic acid synthesis screening (via designing "synthetic homologs" encoded by non-standard DNA sequences)
- Enabling resistance to countermeasures (vaccines, anti-virals)
- Evading biosurveillance systems for detecting disease outbreaks

Misaligned AI could further weaken global biosecurity situation (ferment discord and lack of trust)

Actions to Reduce AIxBio Risks – Primary Prevention

- Recognise the global **metacrisis/polycrisis** [Lawrence et al 2024]: Need fundamental system solutions for: catastrophic risks, climate disruption, conflicts/trade wars, inequities etc

- **International treaties** around advanced AI (for governance & guardrails with verification & enforcement mechanisms)

- Upgrade and strengthen the **Bioweapons Convention** (1972)

- **Whole-of-society approach**: governments, industry, academia, and civil society/philanthropic sector. Helps with tracking evolving AIxBio developments and risks; balancing risks vs potential AI and biotechnology benefits (eg, citizen assemblies/juries)

Primary Prevention – some specifics

- Specifically reduce information hazard around permitting open-weight LLMs
- Tighten laboratory safeguards
- Upgrading DNA synthesis screening systems (eg, using encryption and international networked servers) [Esvelt 2018]
- Potential new regulations & criminal offences [Radcliffe 2025]

For if Prevention Fails: Surveillance, Border Controls & PHSM

- Enhancing **surveillance** & rapid diagnosis (including using AI; metagenomic early-warning systems using sewage [Esvelt 2020])

- **Border controls:** Kill switches [Gervais 2021] for internet connections to avoid concurrent AI-attacks/cyberattacks (eg, islands connected via cables)

- **Border controls:** Attempting exclusion ie, respond quickly to pandemic risk (especially island nations [Boyd et al])

- **Eliminate or mitigate any spread:** State-of-the art public health & social measures (PHSM)

Elimination or Mitigation of Extreme Pandemics

- If attempting elimination: **Stay-at-home requirements** eg, identify ***really*** essential workers & have PPE for all of them (ie, workers in: food supply, grid functioning (electricity, internet, water, sewerage), police/military [Geneva Centre for Security Policy 2023]
- If mitigation: State-of-the art **public health & social measures** (with legal frameworks and high quality communication)
- If mitigation: Capacity to develop & distribute new **therapeutics and vaccines**

Conclusions

- Expert agreement on there being potential **catastrophic risks** from AIxBio (but high uncertainty)

- **Risks may be increasing** with advancements in AI, biotechnology & robotics

- Need to address the global **metacrisis/polycrisis** but also need specific international action for **primary prevention** of AIxBio risks (treaties, governance etc)

- Need to prepare for if prevention fails – so need **enhanced surveillance, border controls and PHSM**

References

- AI Expert Statement (350+ experts): Roose K. A.I. Poses ‘Risk of Extinction,’ Industry Leaders Warn. New York Times 2023;(30 May). https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html.
- Anthropic Report 2025: https://www.anthropic.com/research/agentic-misalignment
- Boyd et al. Impact of Covid-19 Control Strategies on Health and GDP Growth Outcomes in 193 Sovereign Jurisdictions. medRxiv 2025:2025.04.08.25325452. https://www.medrxiv.org/content/medrxiv/early/2025/04/10/2025.04.08.25325452.full.pdf
- Elders 2023: The Elders statement: https://theelders.org/news/elders-urge-global-co-operation-manage-risks-and-share-benefits-ai
- Esvelt 2018: https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1007286
- Esvelt 2020: https://www.effectivealtruism.org/articles/kevin-esvelt-mitigating-catastrophic-biorisks?
- FRI 2025: Forecasting Research Institute (FRI) – Bridget Williams et al. FRI’s white paper: "Forecasting biosecurity risks from large language models" (July 1, 2025) https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/68812b62e85b2808f0366c41/1753295738891/ai-enabled-biorisk.pdf
- Future of Life Institute 2025: https://futureoflife.org/wp-content/uploads/2025/07/FLI-AI-Safety-Index-Report-Summer-2025.pdf
- Geneva Centre for Security Policy 2023; Gopal et al: https://www.gcsp.ch/publications/securing-civilisation-against-catastrophic-pandemics
- Gervais 2021: https://pmc.ncbi.nlm.nih.gov/articles/PMC8576463/
- Karger et al 2023: Forecasting Existential Risks: Evidence from a Long-Run Forecasting Tournament (FRI Working Paper #1), Forecasting Research Institute. https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64abffe3f024747dd0e38d71/1688993798938/XPT.pdf.
- Lawrence et al 2024: https://www.cambridge.org/core/journals/global-sustainability/article/global-polycrisis-the-causal-mechanisms-of-crisis-entanglement/06F0F8F3B993A221971151E3CB054B5E
- Nature 2025: Editorial 17 July 2025: https://www.nature.com/articles/d41586-025-02271-w
- NTI 2025: Nuclear Threat Initiative (statement) https://www.nti.org/analysis/articles/statement-on-biosecurity-risks-at-the-convergence-of-ai-and-the-life-sciences/ [July 2025]
- Radcliffe 2025: Assessing the Accelerated Threat of Bioterrorism in the Age of AI, 49 Wm. & Mary Env’t L. & Pol'y Rev. 763 (2025), https://scholarship.law.wm.edu/wmelpr/vol49/iss3/10
- RAND 2025: Vermeer et al. On the extinction risk from artificial intelligence, RAND: https://www.rand.org/pubs/research_reports/RRA3034-3031.html.
- Sandberg & Nelson 2020: https://www.liebertpub.com/doi/10.1089/hs.2019.0115