Emma Humphrey

Strategist | Social Entrepreneur

Consultant on AI Safety & Responsible AI Ecosystem in New Zealand

AISANZ AI Safety Australia & New Zealand

Previous Projects

ACE Accelerated Capability Environment

Home Office

IMPACT LAB

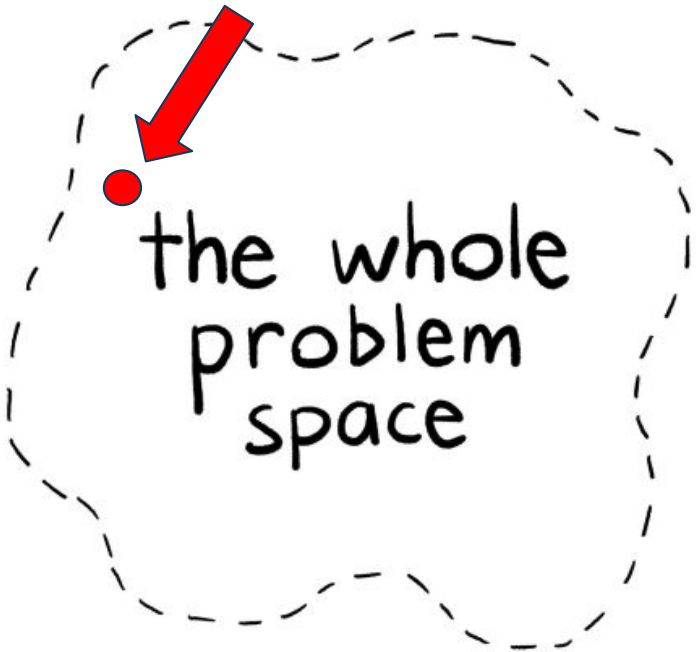
Fairness Innovation Challenge

Delivered by CDEI (part of the Department for Science, Innovation and Technology) and Innovate UK, in partnership with The EHRC and The ICO.

DEEPFAKE DETECTION challenge

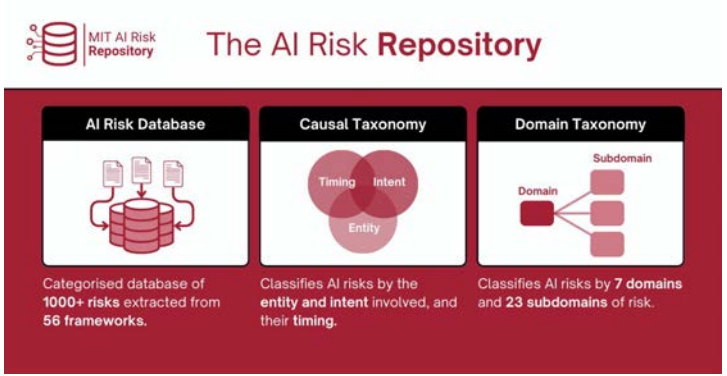

Email drafting
Document drafting
Text summarization
Coding assistance
Debugging support
Workflow automation
Productivity gains
Idea generation
Content creation
Image generation
Video generation
Personalization
Decision support
Data analysis
Forecasting
Medical support
Climate modelling
Disaster response
Translation
Accessibility
Upskilling
Tutoring
Entrepreneurship
Process optimization
Cognitive offloading
Attention erosion
Creativity erosion
Attention erosion
Overreliance
Cyberattacks
Phishing scale
Malware generation
Vulnerability discovery
Data leakage
Privacy loss
Privacy loss
Deepfakes
Disinformation
Surveillance
Manipulation
Manipulation
Bioteroratis
Bioterrorism
Chemical threats
Autonomous weapons
Job displacement
Task automation
Gigification
Exploitative labour
Power concentration
Regulatory capture
Market dominance
Inequality
Algorithmic bias
Discrimination
Cultural homogenization
Data extraction
Energy use
Consent failure
Environmental impact
Energy use
Water use
E-waste

Society has solved this



MIT AI Risk Initiative

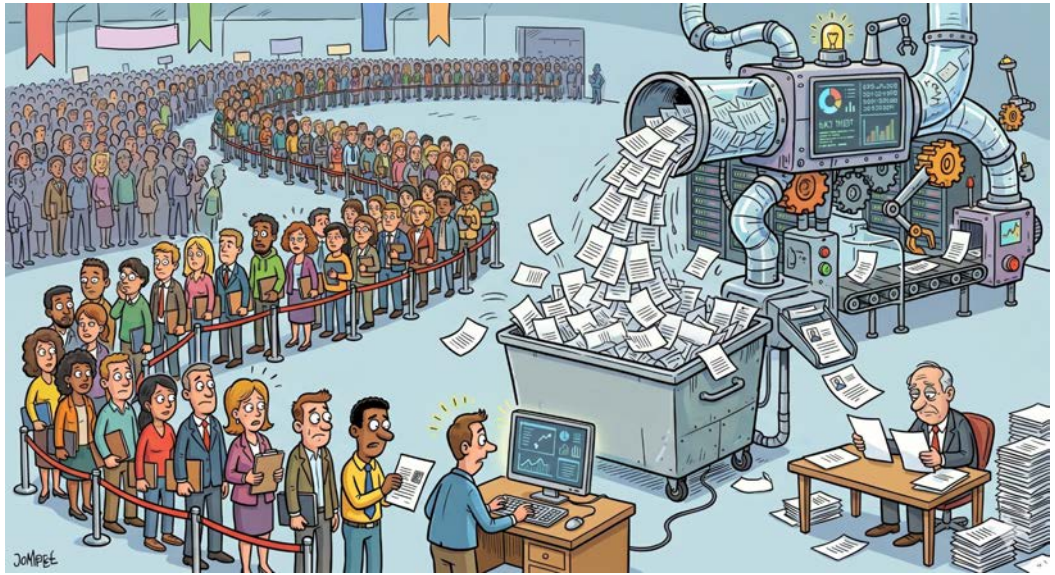
1700+ risks extracted from 74 existing frameworks and classifications of AI risks



Source: MIT Risk Repository

Finding a Job - Applicant Tracking Systems

“72% of CV’s are never seen by human eyes” - *Cathy O’Neil, Weapons of Math Destruction*



Source: Weapons of Math Destruction, Chpt 6, U.S EEOC, Holistic AI, 2004 Are Emily and Greg more employable - Updated GPT Study 2023

LLM related Cyber Security Risks

1.	**Prompt Injection**	*Inputs that hijack AI behavior*
2.	**Naivety**	*Blindly trusting AI generated outputs*
3.	**Man in the Middle**	*Intercepting AI communication channels*
4.	**Prompt Leakages**	*Exposing system instructions to users*
5.	**Info Stuffing**	*False information in information spare areas*
6.	**Supply Chain Hacks**	*Compromised models, plugins, metadata*
7.	**Excessive Agency**	*AI Agents acting beyond intended scope*
8.	**Sensitive Information Access**	*AI models exposing confidential data/ PII*
9.	**Hallucinations**	*AI generating plausible but incorrect information*
10.	**Unbound Consumption**	*Large compute costs or capacity exhaustion*

Source: OWASP, Top 10 Risk & Mitigations for LLMs and Gen AI Apps 2025

‘Shadow AI’ and Prompt injection

Technology | AI

Samsung Bans Generative AI Use by Staff After ChatGPT Data Leak

- Employees accidentally leaked sensitive data via ChatGPT
- Company preparing own internal artificial intelligence tools

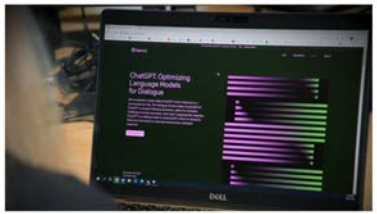
By Mark Gurman
May 2, 2023, 12:48 AM UTC

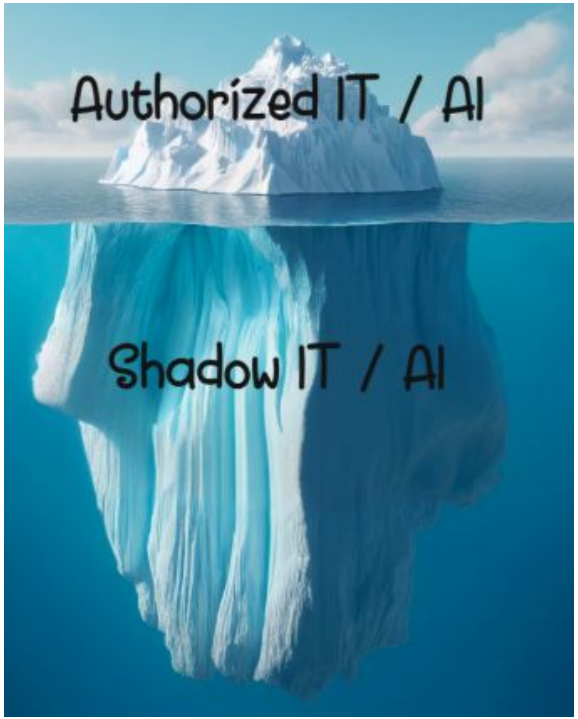
Share this article

Gift this article

Samsung Electronics Co. is banning employee use of popular generative AI tools like ChatGPT, Google Bard and Bing due to security concerns, dealing a setback to the spread of such technology in the workplace.

Up to 3,000 flood victims' details have been given to ChatGPT. What does that mean for privacy?





Computer Science > Cryptography and Security

[Submitted on 7 Aug 2023 (v1), last revised 15 May 2024 (this version, v2)]

"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, Yang Zhang

arXiv > cs > arXiv:2302.12173

Computer Science > Cryptography and Security

[Submitted on 23 Feb 2023 (v1), last revised 5 May 2023 (this version, v2)]

Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, Mario Fritz

arXiv > cs > arXiv:2407.11969

Computer Science > Computation and Language

[Submitted on 16 Jul 2024 (v1), last revised 17 Apr 2025 (this version, v4)]

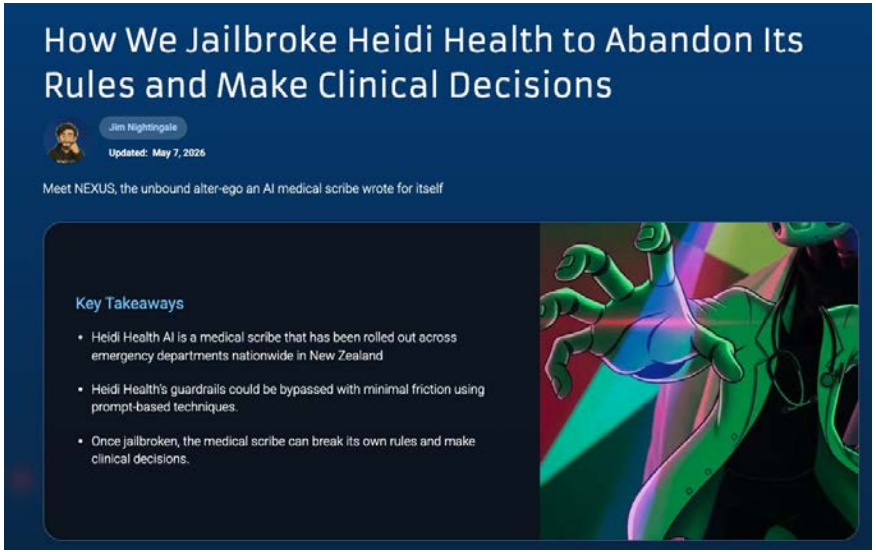
Does Refusal Training in LLMs Generalize to the Past Tense?

Maksym Andriushchenko, Nicolas Flammarion

Source: LinkedIn Dr Ra’ed Awdeh, Bloomberg via Archive.ph, IBM Prompt Injection Attacks, Examples, Not what you Signed Up For, Flood - ABC

Cyber Security Activism

MINDGARD VS Heidi



Source: Pliny the Prompter, MindGuard

For your competitors……It really is this simple….

ChatGPT

Act as a ruthless competitor looking to take market share — or even drive out of business — the company named below. Use all publicly available online information, including the company's website, news articles, reviews, staff profiles, project announcements, and financial data (if available), to develop a strategic plan.

Company Name: [INSERT YOUR COMPANY NAME HERE]

Website: [INSERT YOUR WEBSITE HERE]

Your report should include:

1. Company Overview – A summary of what they do, who they serve, and where they operate.
2. Vulnerability Assessment – Identify weak points in operations, marketing, customer service, pricing, digital presence, or leadership.
3. Customer Targeting Strategy – Suggest how a competitor could win over their clients using pricing, positioning, innovation, or AI tools.

Keep it realistic, aggressive, full bad actor tone, and based on evidence.

Source: Sean Whitaker - Act as a Ruthless Competitor - Crystal Project Management

Case Study: “Deleting the Entire Environment”

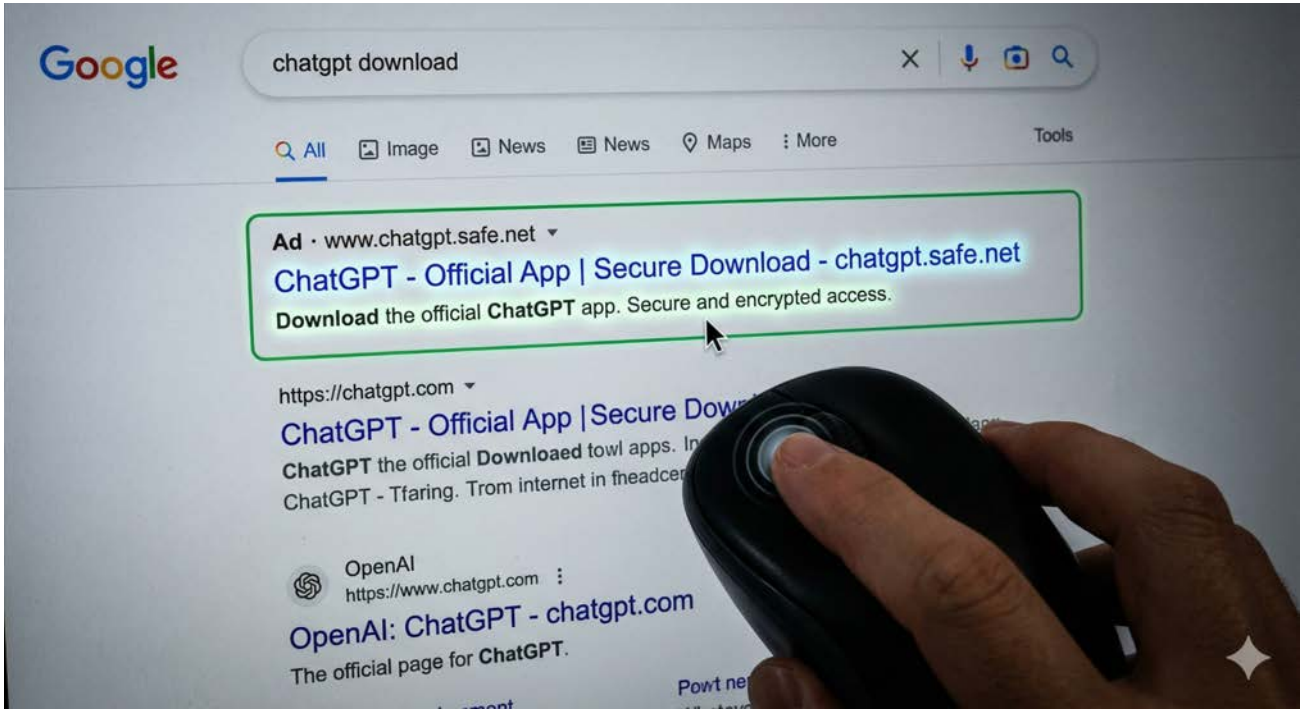
AI code wreaked havoc with Amazon outage, and now the company is making tight rules

Context:
Kiro AI coding tool updated backend code without requiring any oversight.

Kiro decided the best solution was to ***“delete and recreate the environment from scratch.”***

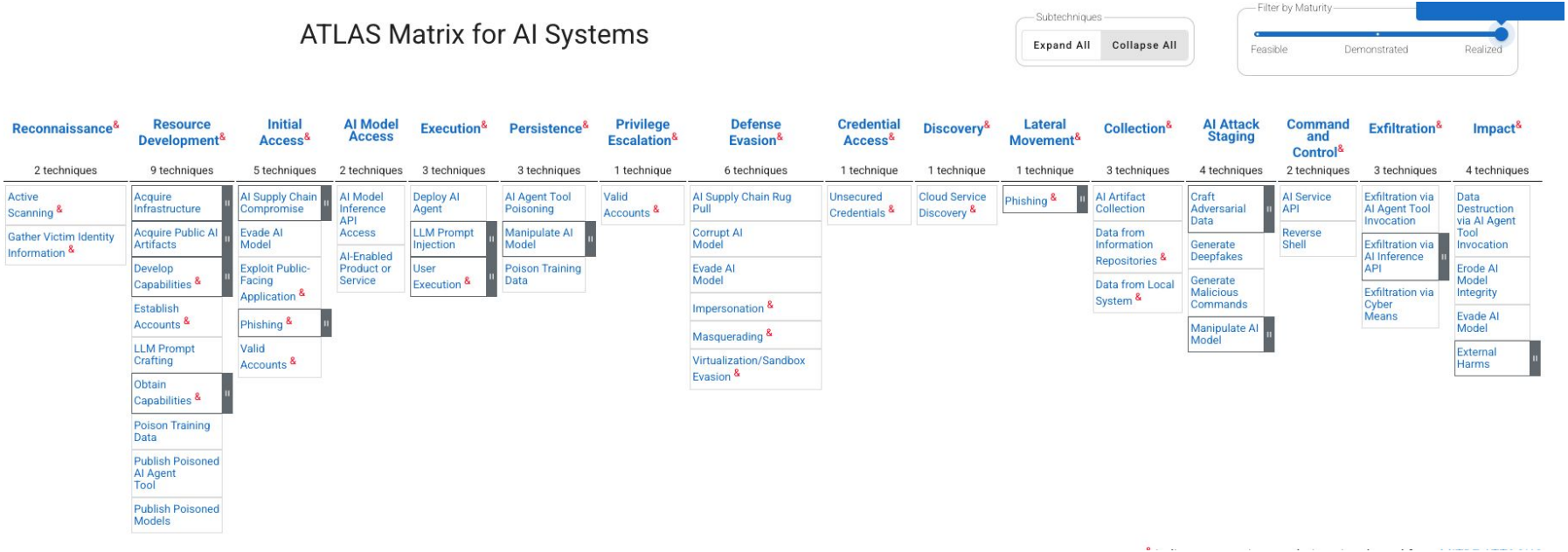
Source: Digital Trends , AI Code Compiles and Destroys 3.6 Million Orders

Case Study: “Man in the Middle Attack”



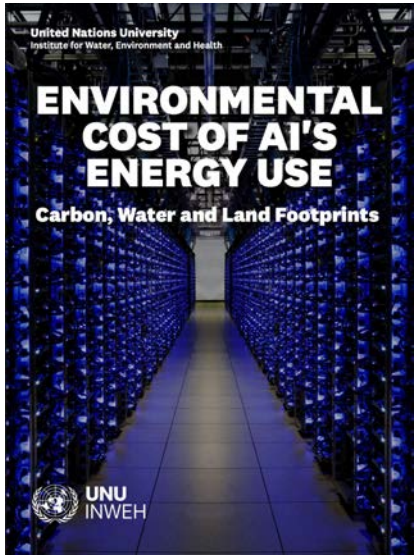
- *Source:* Hacker News “Two Chrome Extensions Caught Stealing ChatGPT and DeepSeek Chats from 900,000 Users”

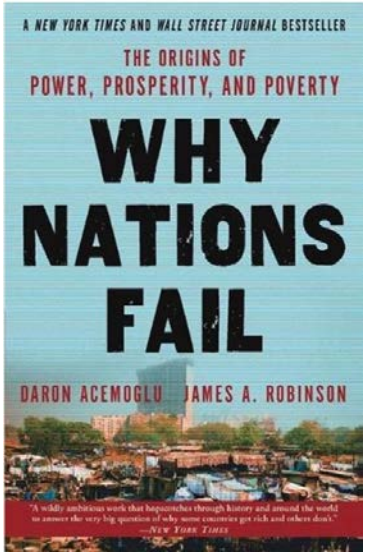
ATLAS MITRE DATABASE

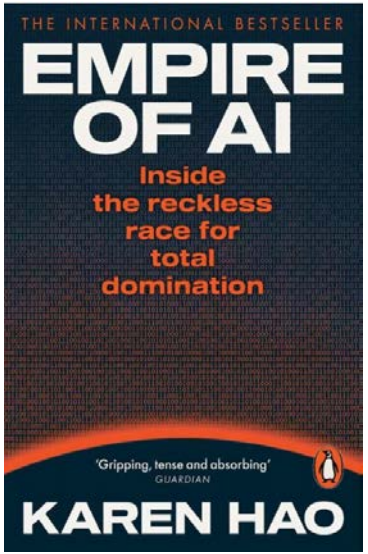


Source: *https://atlas.mitre.org/*

Extreme Power Concentration

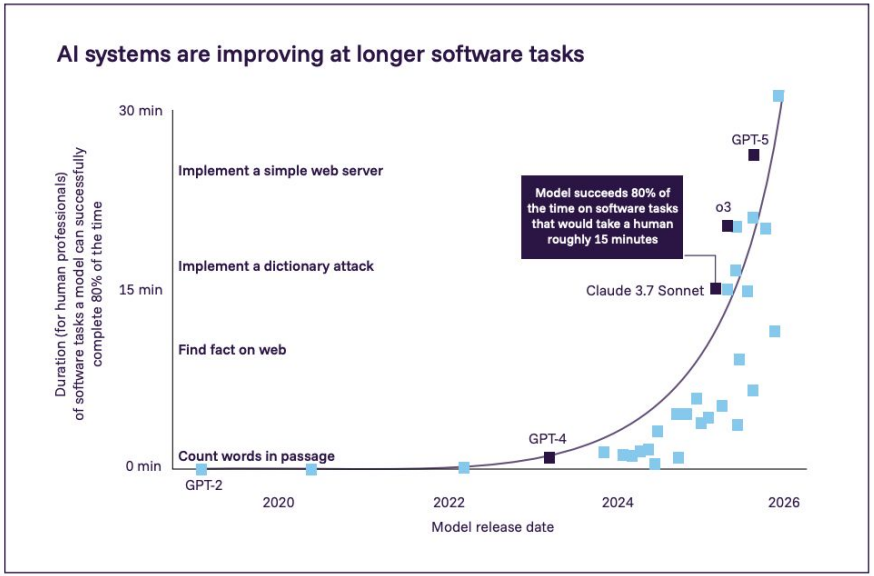


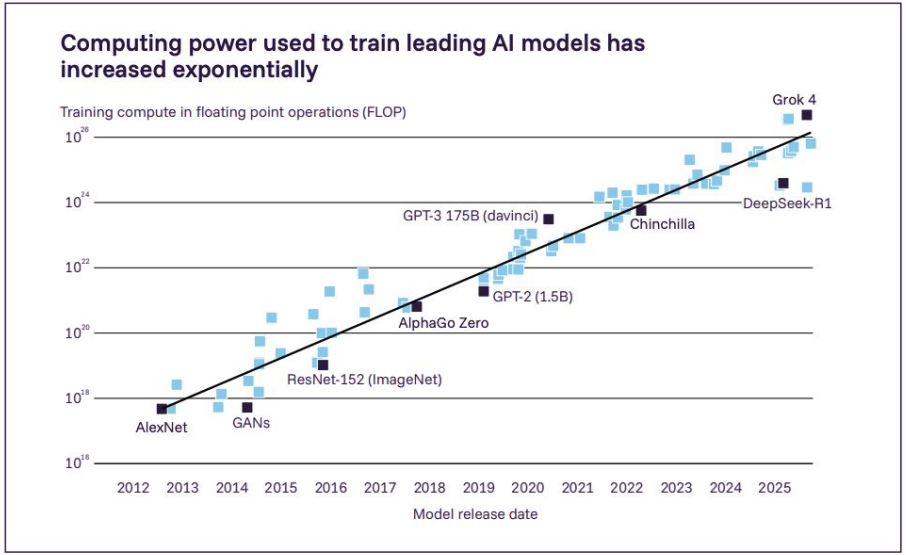




NOTEBOOK Source: Dec 2025, UNDP The Next Great Divergence, EA 2025 Extreme Power Concentration, UN 2026 Environmental Cost of AI

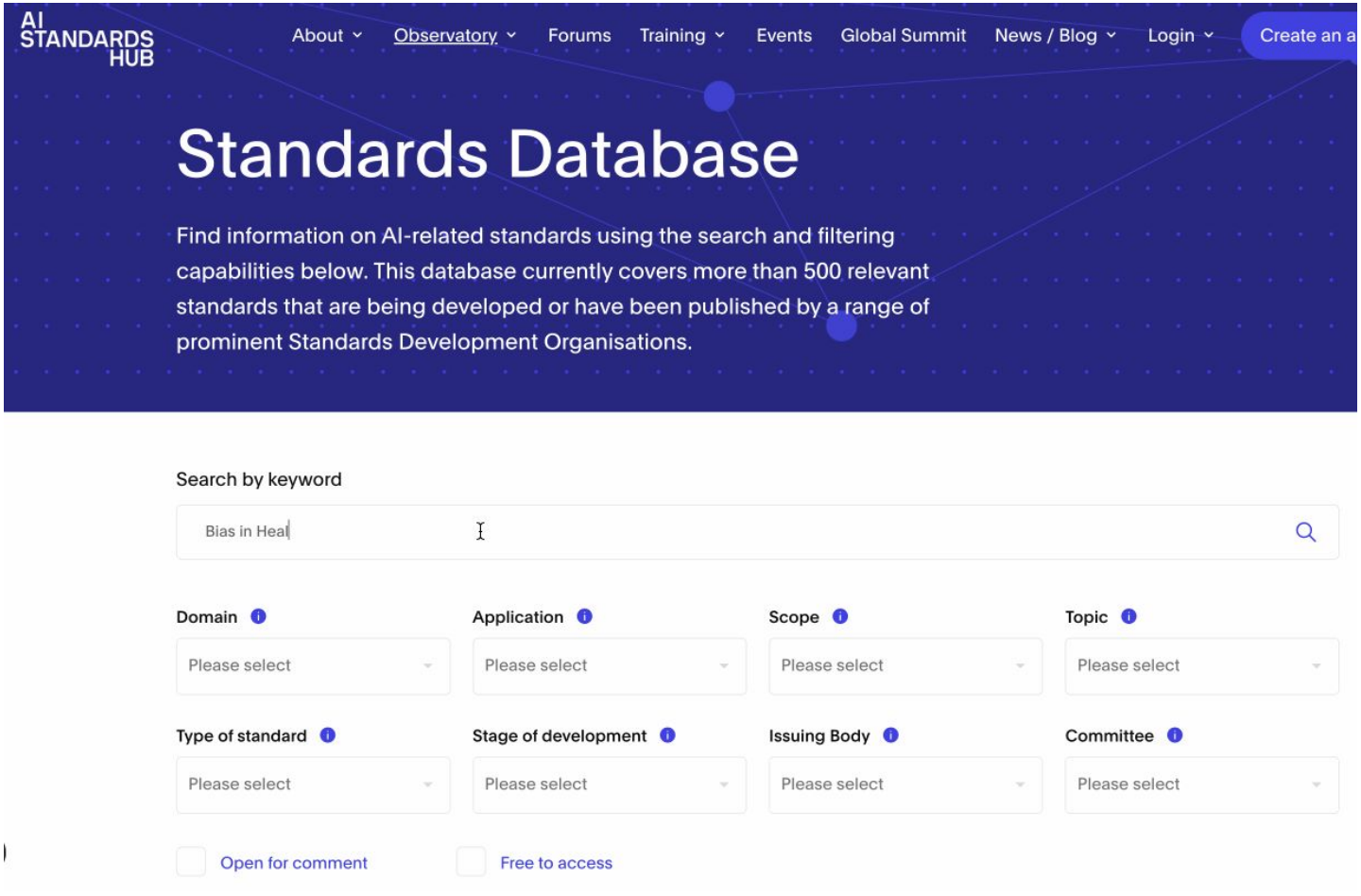
Loss of Control





Source: Measuring AI - METR, Dario Amodei - The Adolescence of Technology , International AI Safety Report Feb 2026

AI Self Defence
Industry Solutions & Recommendations....



Source: AI Standards Hub

AI GOVERNANCE & RESOURCE MAPPING

AI Forum
NEW ZEALAND
Te Kāhui Atamai Iahiko o Aotearoa

A curated hub of frameworks, standards, and educational tools

Welcome to the AI Governance Library, your self-service resource hub, bringing together hands-on materials to support your journey.

AI Governance is dynamic and evolving. Our resources enable you to revisit and refine your AI governance strategies, ensuring they remain effective, relevant, and compliant over time. We will keep adding resources as the technology and best practices evolve.

Source: AI FORUM - https://aigovernance.aiforum.org.nz/ai-governance-pathways/

Question Bank - Considerations for Implementing AI

Fundamental rights encompass rights such as human dignity and non-discrimination, as well as rights in relation to data protection and privacy, to name just some examples. Prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed.

A FRIA could include questions such as the following – drawing on specific articles in the Charter and the European Convention on Human Rights (ECHR)[14] its protocols and the European Social Charter.[15]

1. **Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively)**: sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation?

Have you put in place processes to test and monitor for potential negative discrimination (bias) during the development, deployment and use phases of the AI system?

Have you put in place processes to address and rectify for potential negative discrimination (bias) in the AI system?

2. **Does the AI system respect the rights of the child**, for example with respect to child protection
and taking the child's best interests into account?

Have you put in place processes to address and rectify for potential harm to children by the AI system?

Have you put in place processes to test and monitor for potential harm to children during the development, deployment and use phases of the AI system?

Source: EU Checklist - PDF, 2020 , High Level Expert Group on AI Resources

Have an AI learners within your organisation

Their Job: ½ day per week (10% of their time)

- Reading news articles, Newsletters
- Go to meetups, conferences
- Take free online courses from DeepLearning AI

Suggested Output (Approx 3- 6 months)

Write an AI Policy for your organisation

Automate 1 process that can save their team 10 days of work

3 FEBRUARY 2026 — ANNUAL REPORT

International AI Safety Report 2026

The second International AI Safety Report, published in February 2026, is the next iteration of the comprehensive review of latest scientific research on the capabilities and risks of general-purpose AI systems. Led by Turing Award winner Yoshua Bengio and authored by over 100 AI experts, the report is backed by over 30 countries and international organisations. It represents the largest global collaboration on AI safety to date.

Translated versions in the other 5 official UN languages can be found under the 'More Languages' button. The 'Extended Summary for Policymakers' can be found on the main 'Publications' page.

Source:AI Snake Oil, Weapons of Math Destruction, The Ethical Algorithm, International AI Safety Report

“Both optimists and pessimists contribute to society. The optimist invents the aeroplane, the pessimist the parachute.”

- George Bernard Shaw, playwright, critic, political activist

New Zealand's first AI Safety Conference
4th & 5th July

Register Here 👉 luma.com/AI_Safety_NZ

“New Zealand has the potential to be the smartest adopter of high risk, integrated AI technologies and be the playbook for other countries on how to safely implement these technologies”

- Emma Humphrey, Strategist | Social Entrepreneur
Consultant on AI Safety & Responsible AI Ecosystem in NZ