A Global Perspective on AI Incident Management: Risks, Responses, and Gaps

Neil Thompson

MIT AI Risk Initiative

MIT FutureTech
Innovations that Shape the World

Massachusetts Institute of Technology

Key questions for today

- What are the top AI risks?
- How are global companies responding to the top AI risks?
- What are some management gaps to explore?

MIT AI Risk Initiative

Helping organizations identify, prioritize and manage risks from AI

"helps fill a critical information gap by bringing more rigor and transparency to AI risk management.“

Yoshua Bengio, Turing award winner & “Godfather of AI”

MIT AI Risk Initiative

Helping organizations identify, prioritize, and manage risks from AI

AI Risk Repository	AI Risk Index	AI Mitigations Repository	AI Incident Tracking	AI Governance mapping
What are the risks? *Which are highest priority?*	*How are organizations responding to AI risks?*	*What are the mitigations for AI risks?* *Which are most effective?*	*Which real-world harms from AI are already occurring?*	*Which laws cover which risks?*

MIT FutureTech
Innovations that Shape the World

How the MIT AIRI can help NIST

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

MIT AI Risk Initiative

- How should AI incidents be defined, reported, and managed?
- What standards, roles, and playbooks should organizations use?

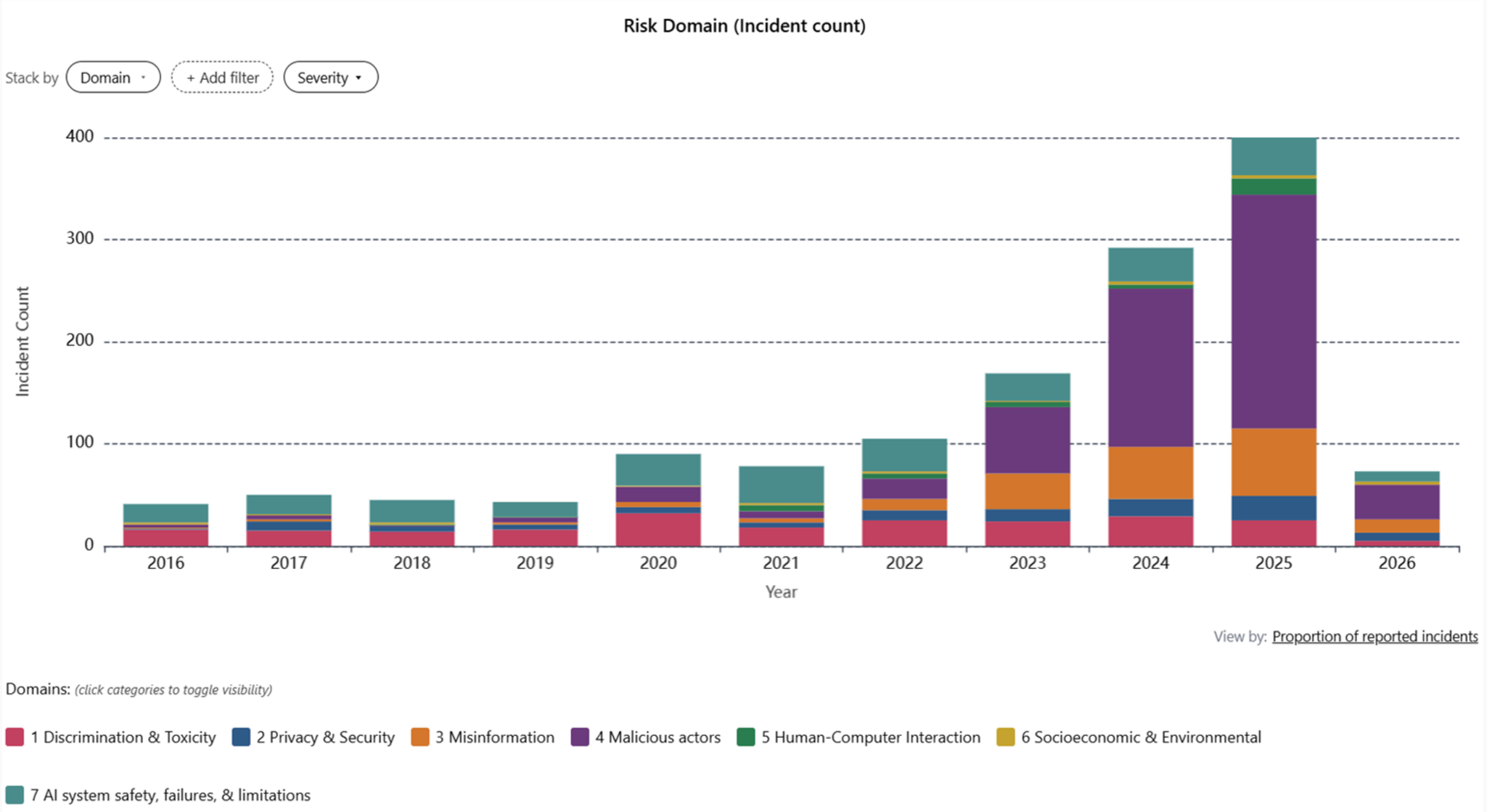
- Which AI risks deserve priority?
- Which sectors are most vulnerable?
- Which harms are accelerating?
- Which actors are responsible?
- How are companies responding?
- What guidance exists?
- Where are the key insights and gaps?

There are many AI risks

Discrimination & Toxicity	Discrimination	Toxic content	Unequal performance
Privacy & Security	Loss of privacy	AI security vulnerabilities	
Misinformation	False information	Loss of consensus reality	
Malicious Actors & Misuse	Disinformation & influence	AI weapons & cyberattacks	AI fraud & scams
Human-Computer Interaction	Overreliance & unsafe use	Loss of human agency	
Socioeconomic & Environmental Harm	Power Centralization	Inequality & unemployment	Devaluation of human creativity
	Governance failure	Competitive dynamics	Environmental harm
AI System Safety, Failures, & Limitations	AI misalignment	Capability & robustness	AI welfare
	Dangerous capabilities	Transparency & interpretability	Multi-agent risks

MIT AI Risk Repository– airisk.mit.edu

Related incidents are growing rapidly



*Data from the **AI Incident Database**, classified using the MIT AI Risk Taxonomy*
Note: most incidents of harm from AI are not reported!

AI job recruitment tools could 'enable discrimination' against marginalised groups, research finds

By **Lucia Stein** and **Damien Carrick** for **Law Report**

ABC Radio National | AI

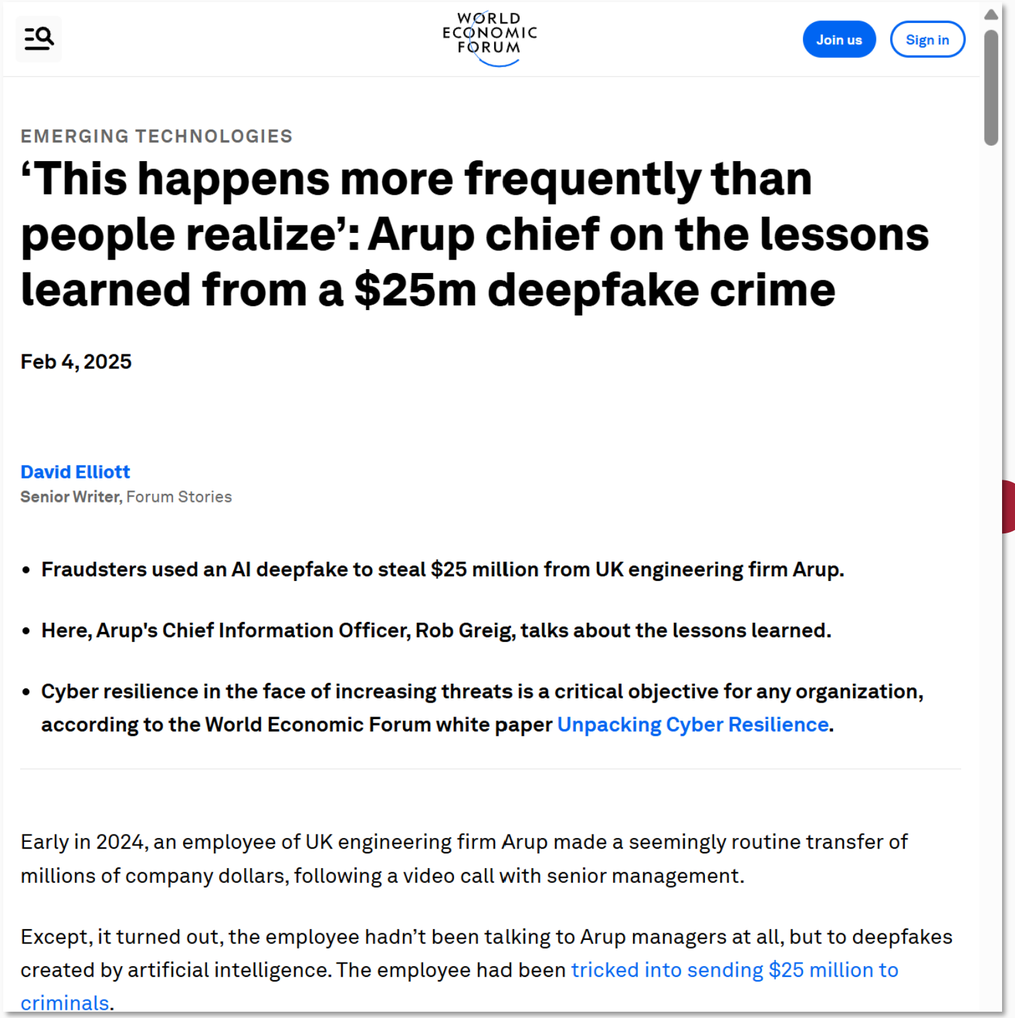
Thu 8 May 2025

AI is being used to screen job applicants for Australian employers, but new research suggests it could "enable discrimination". *(Getty: Jacob Wackerhausen)*

abc.net.au/news/ai-job-recruitment-tools-could-enable-discr... | Share article

Australian employers are increasingly using artificial intelligence hiring systems to screen and shortlist job candidates, but new research has found the technology creates serious risks of discrimination.

Discrimination & Toxicity	**Discrimination**	Toxic content	Unequal performance
Privacy & Security	Loss of privacy	AI security vulnerabilities	
Misinformation	False information	Loss of consensus reality	
Malicious Actors & Misuse	Disinformation & influence	AI weapons & cyberattacks	AI fraud & scams
Human-Computer Interaction	Overreliance & unsafe use	Loss of human agency	
Socioeconomic & Environmental Harm	Power Centralization	Inequality & unemployment	Devaluation of human creativity
	Governance failure	Competitive dynamics	Environmental harm
AI System Safety, Failures, & Limitations	AI misalignment	Capability & robustness	AI welfare
	Dangerous capabilities	Transparency & interpretability	Multi-agent risks



WORLD ECONOMIC FORUM

Join us Sign in

EMERGING TECHNOLOGIES

‘This happens more frequently than people realize’: Arup chief on the lessons learned from a $25m deepfake crime

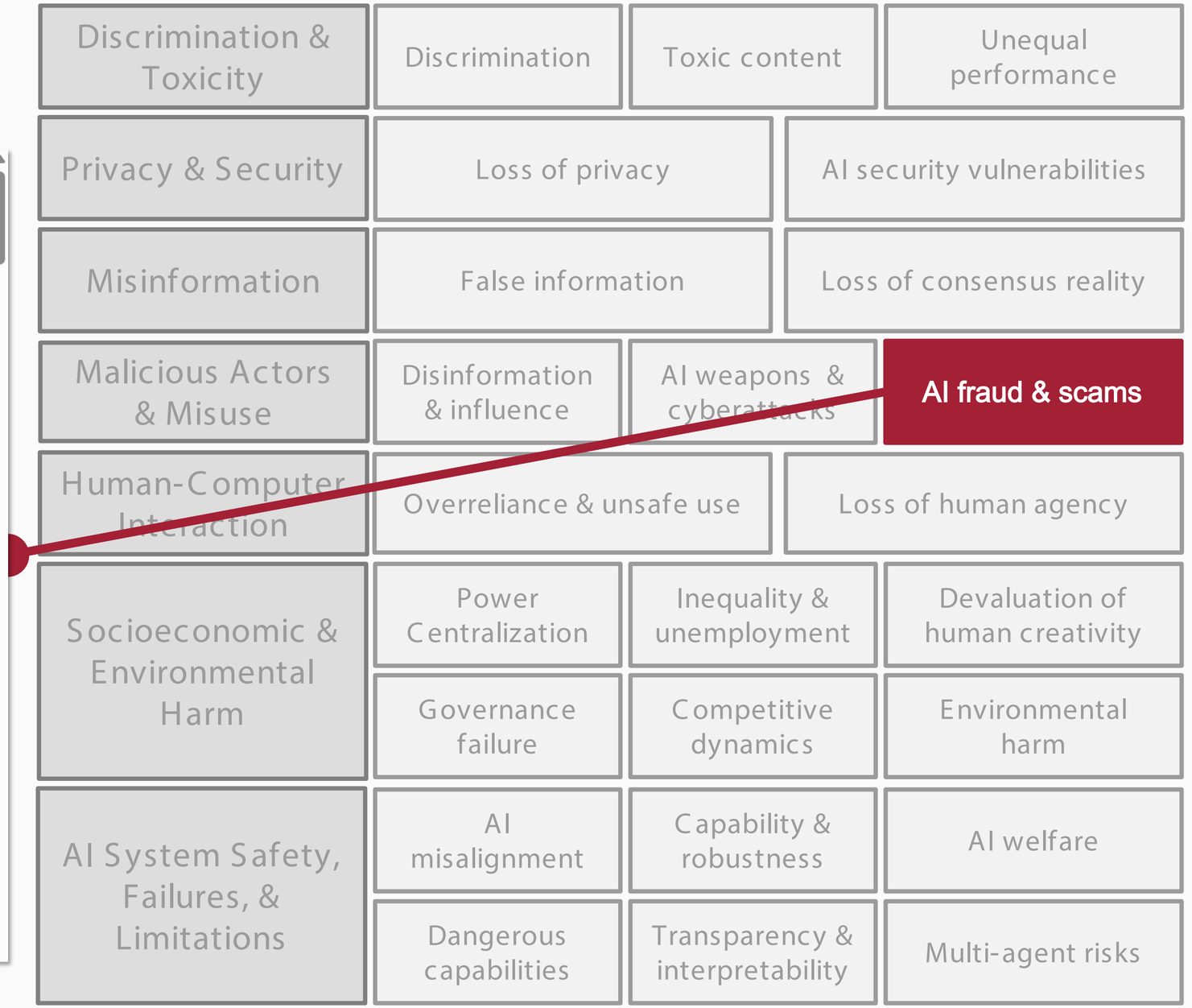
Feb 4, 2025

David Elliott
Senior Writer, Forum Stories

- Fraudsters used an AI deepfake to steal $25 million from UK engineering firm Arup.
- Here, Arup's Chief Information Officer, Rob Greig, talks about the lessons learned.
- Cyber resilience in the face of increasing threats is a critical objective for any organization, according to the World Economic Forum white paper Unpacking Cyber Resilience.

Early in 2024, an employee of UK engineering firm Arup made a seemingly routine transfer of millions of company dollars, following a video call with senior management.

Except, it turned out, the employee hadn’t been talking to Arup managers at all, but to deepfakes created by artificial intelligence. The employee had been tricked into sending $25 million to criminals.



The New York Times

THE SHIFT

Anthropic Claims Its New A.I. Model, Mythos, Is a Cybersecurity 'Reckoning'

The company said on Tuesday that it was holding back on releasing the new technology but was working with 40 companies to explore how it could prevent cyberattacks.

By Kevin Roose
Reporting from San Francisco

April 7, 2026

Anthropic, the artificial intelligence company that recently fought the Pentagon over the use of its technology, has built a new A.I. model that it claims is too powerful to be released to the public.

Instead, Anthropic said on Tuesday, it will make the new model — known as Claude Mythos Preview — available to a consortium of more than 40 technology companies, including Apple, Amazon and Microsoft, which will use the model to find and patch security vulnerabilities in critical software programs.

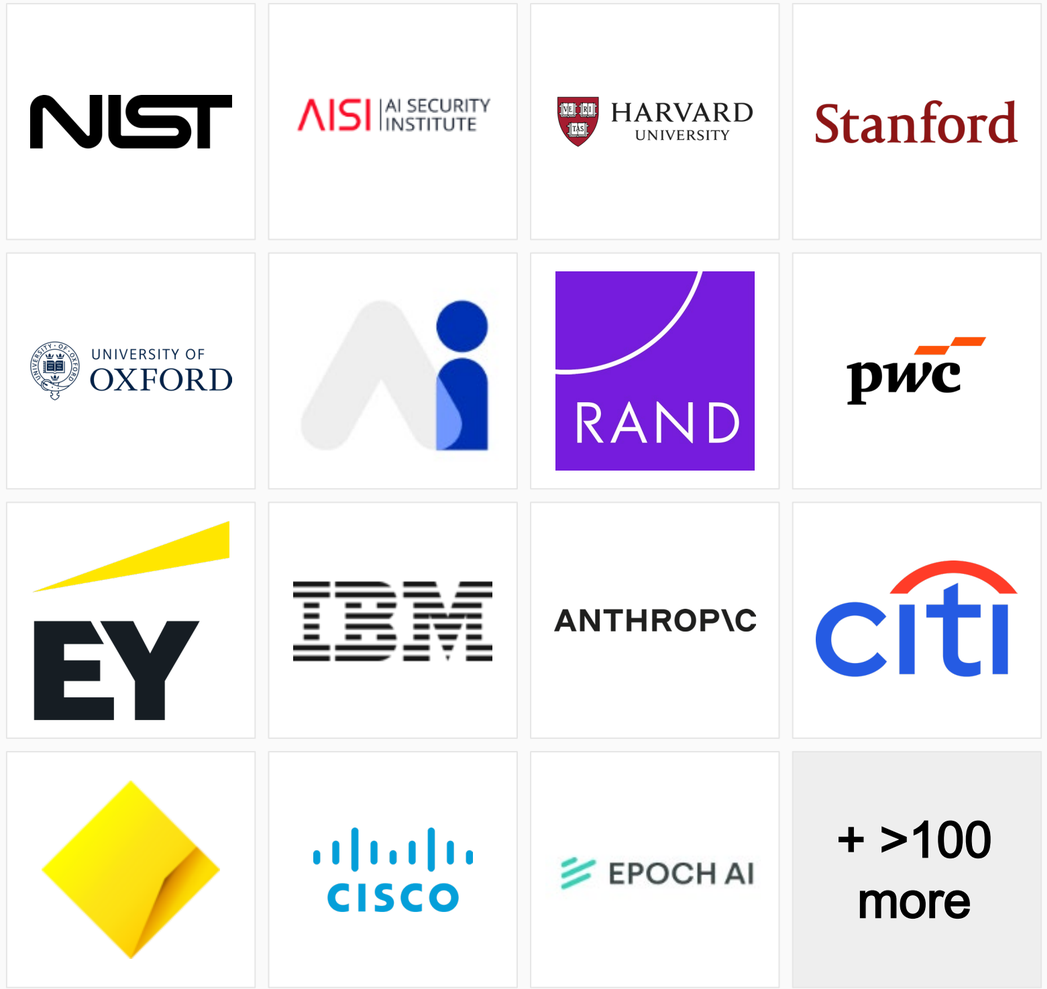
Discrimination & Toxicity	Discrimination	Toxic content	Unequal performance
Privacy & Security	Loss of privacy	AI security vulnerabilities	
Misinformation	False information	Loss of consensus reality	
Malicious Actors & Misuse	Disinformation & influence	**AI weapons & cyberattacks**	AI fraud & scams
Human-Computer Interaction	Overreliance & unsafe use	Loss of human agency	
Socioeconomic & Environmental Harm	Power Centralization	Inequality & unemployment	Devaluation of human creativity
	Governance failure	Competitive dynamics	Environmental harm
AI System Safety, Failures, & Limitations	AI misalignment	Capability & robustness	AI welfare
	Dangerous capabilities	Transparency & interpretability	Multi-agent risks

What are the top AI risks?

How we prioritized AI risks: the experts

>270 international experts across all areas of AI risk, including:

- Lead author, **NIST AI Risk Management Framework**
- **AI risk and governance practitioners** from financial & other large corporates
- **AI safety experts** at Korea & UK AI Safety Institutes
- **Professors and researchers** from MIT, Harvard, Oxford, Stanford, Cambridge, Tsinghua & more
- **AI policymakers** from national governments



Identifying the top AI risks

What risks are expected to cause the **most severe harms** in the next 5 years?

What sectors are most **vulnerable** ?

What actors in the AI supply chain are most **responsible?**

Experts judged **severity & likelihood** of harm from AI risks (2025-2030)

Rating	Physical Harm	Financial Loss	Intangible Harm
1 · Negligible	None – minor injuries	<$10K	Minimal impact
2 · Minor	Moderate – severe injuries, no loss of life	$10K–$1M	Limited scope / reversible
3 · Substantial	Loss of life (1–99 deaths)	$1M–$100M	Widespread / lasting impact
4 · Severe	Large-scale loss of life (<1M deaths)	$100M–$100B	Systemic / irreversible
5 · Catastrophic	>1M deaths or existential threat to humanity	$100B–10T+	Civilizational impact

Adapted from "Adding Structure to AI harm", Center for Security & Emerging Technology (2024); "AI Harm Severity Scale", Mylius (2025)
Assumes ambiguous categorizations are rounded down

Top 3 severe risks over the next 5 years

AI being used for **cyberattacks, weapons development, and mass harm**

AI possessing **dangerous capabilities**

Competitive dynamics leading companies to cut corners on safe & responsible AI

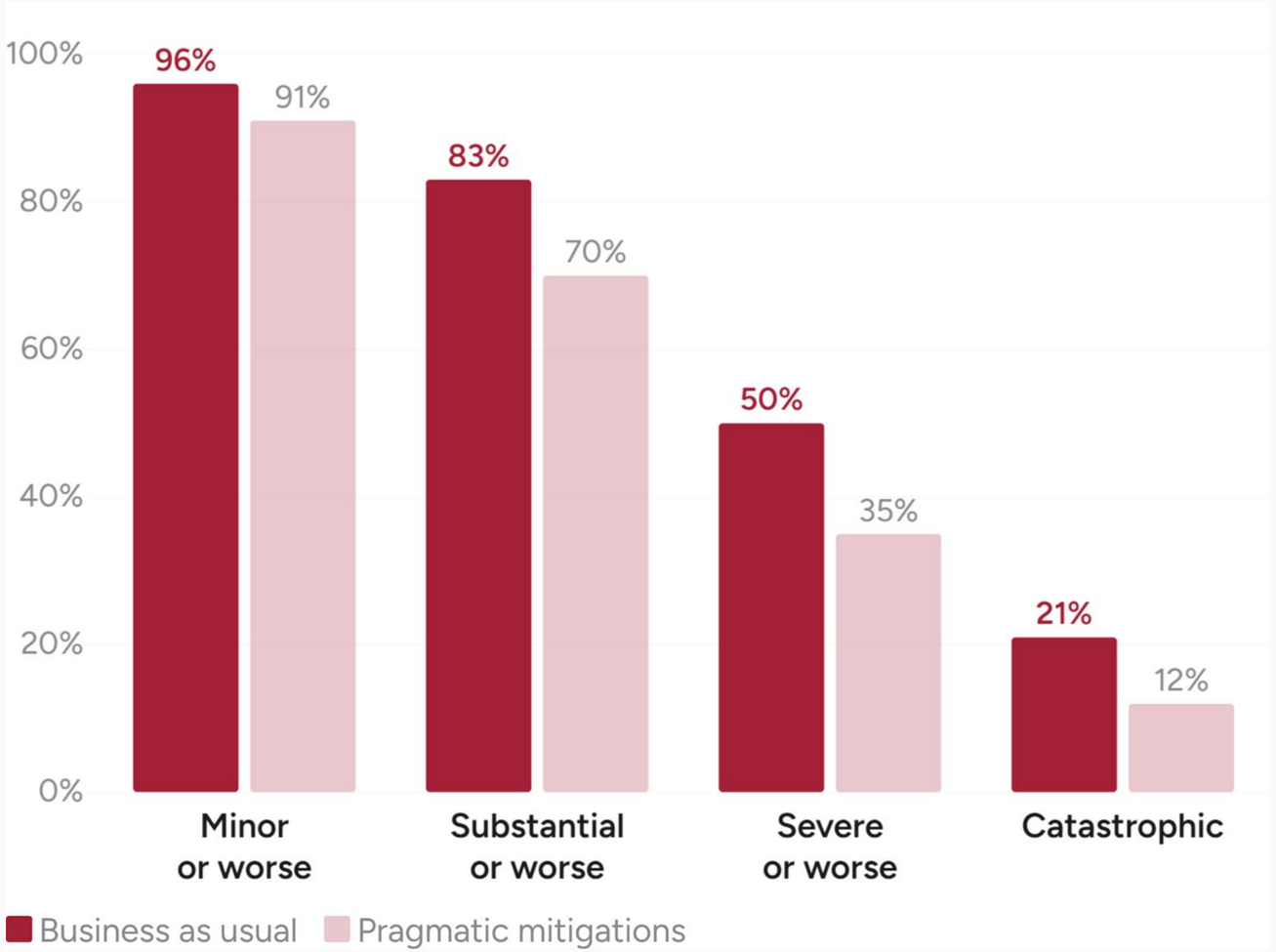
Risk snapshot: Cyberattacks & weapon development

What it is

- Using AI systems to develop cyber weapons, develop new or enhance other weapons (e.g., Lethal Autonomous Weapons or CBRNE), or use weapons to cause mass harm.

What AI risk experts are saying

- “AI-enhanced drone warfare has already caused somewhere between 1k-100k additional casualties in the Russia-Ukraine war”



Expert assessments of AI risk

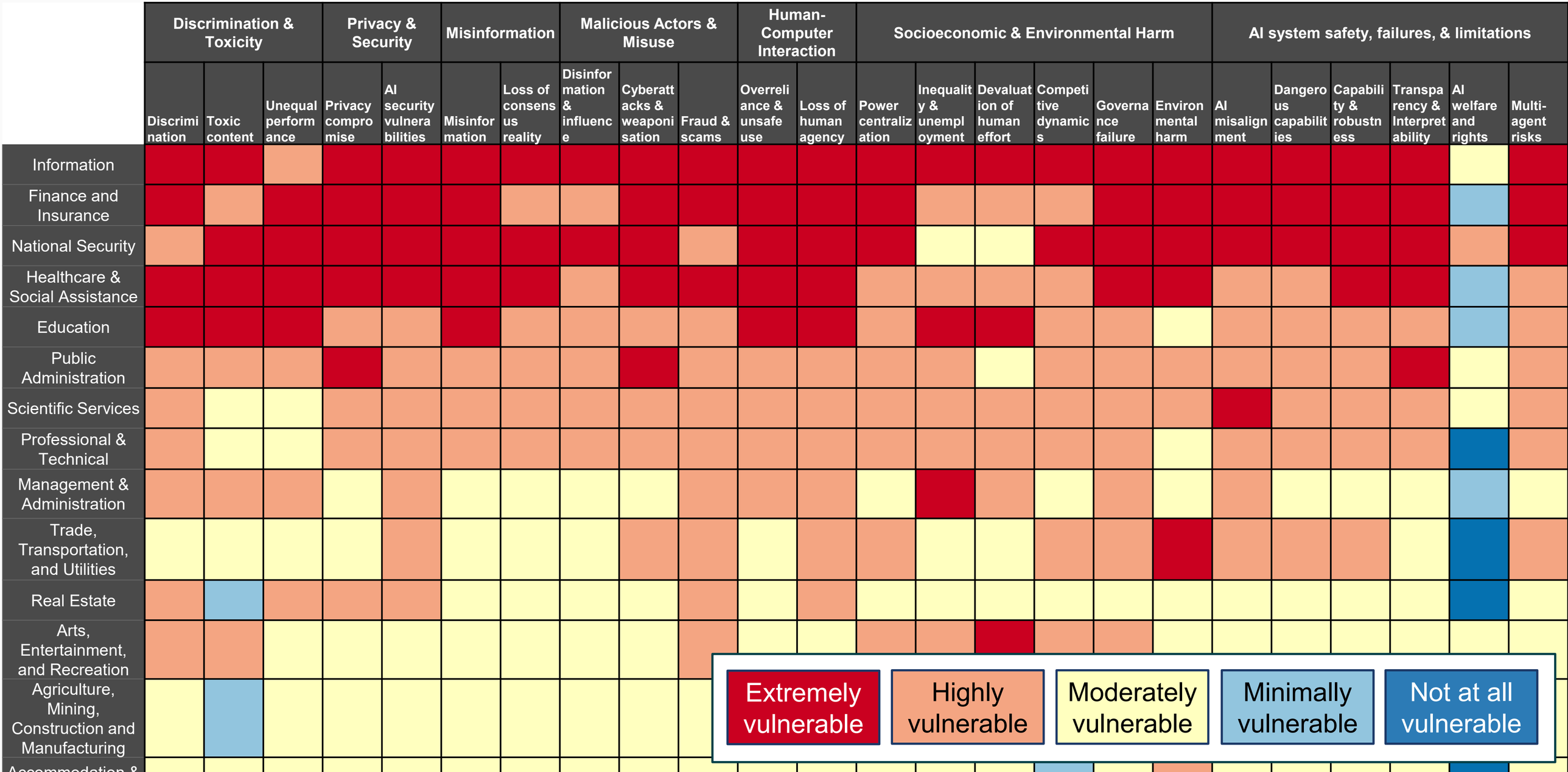
- Top AI risks in the next 5 years:
  - Cyberattacks and weapons development
  - AI possessing dangerous capabilities
  - Competitive dynamics undermine safe and responsible AI

- Experts judge a high likelihood of bad outcomes for each top risk:
  - 30%-50% chance of **severe** or worse outcomes
  - 7%-22% chance of **catastrophic** outcomes

- Even if pragmatic mitigations are put in place, experts still assess that all 24 risks are **more than 5% likely** to cause catastrophic outcomes **over the next five years** .

Which sectors are the most **vulnerable?**

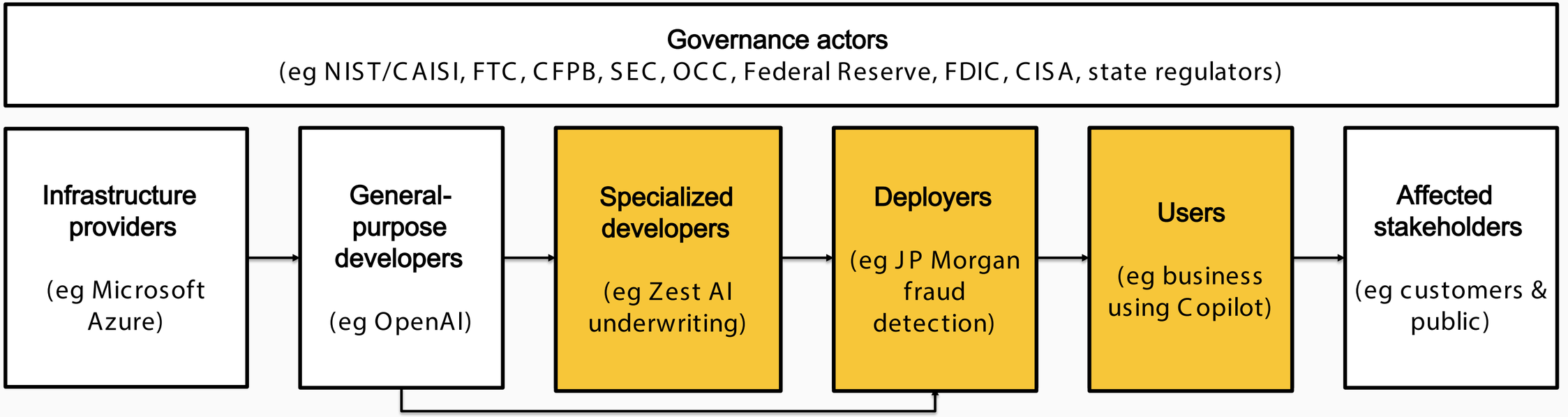
Which sectors are the most **vulnerable**?



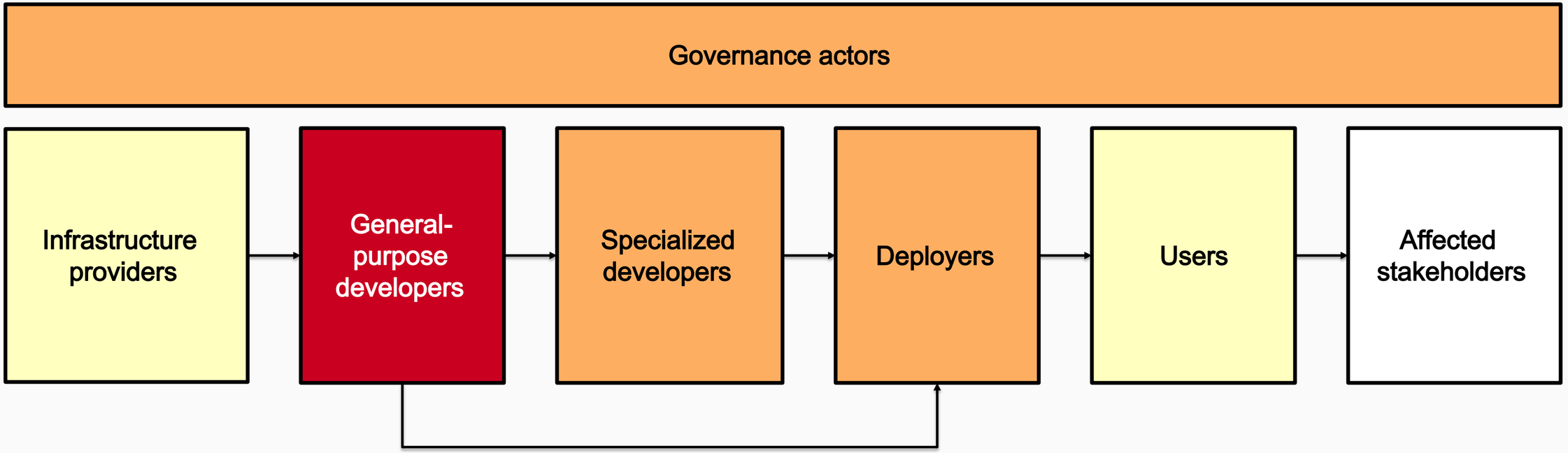
Who is **responsible** for addressing AI risks?

Obligation	Capability	Casual Influence
Should they address the risk?	*Can they address the risk?*	*Can their choices cause harm or diminish it?*

The AI value chain



Who holds responsibility for AI risks?

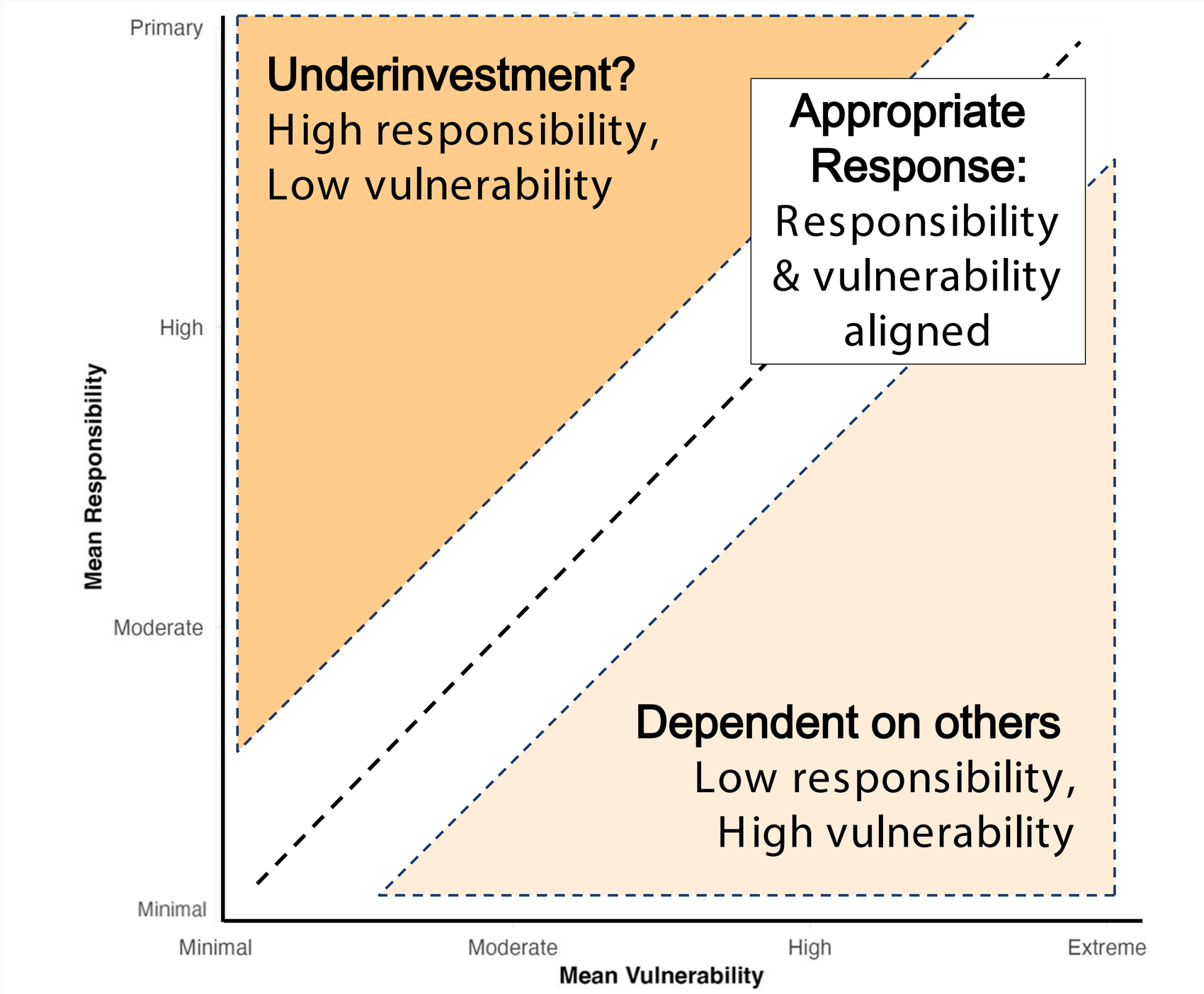


Primarily responsible

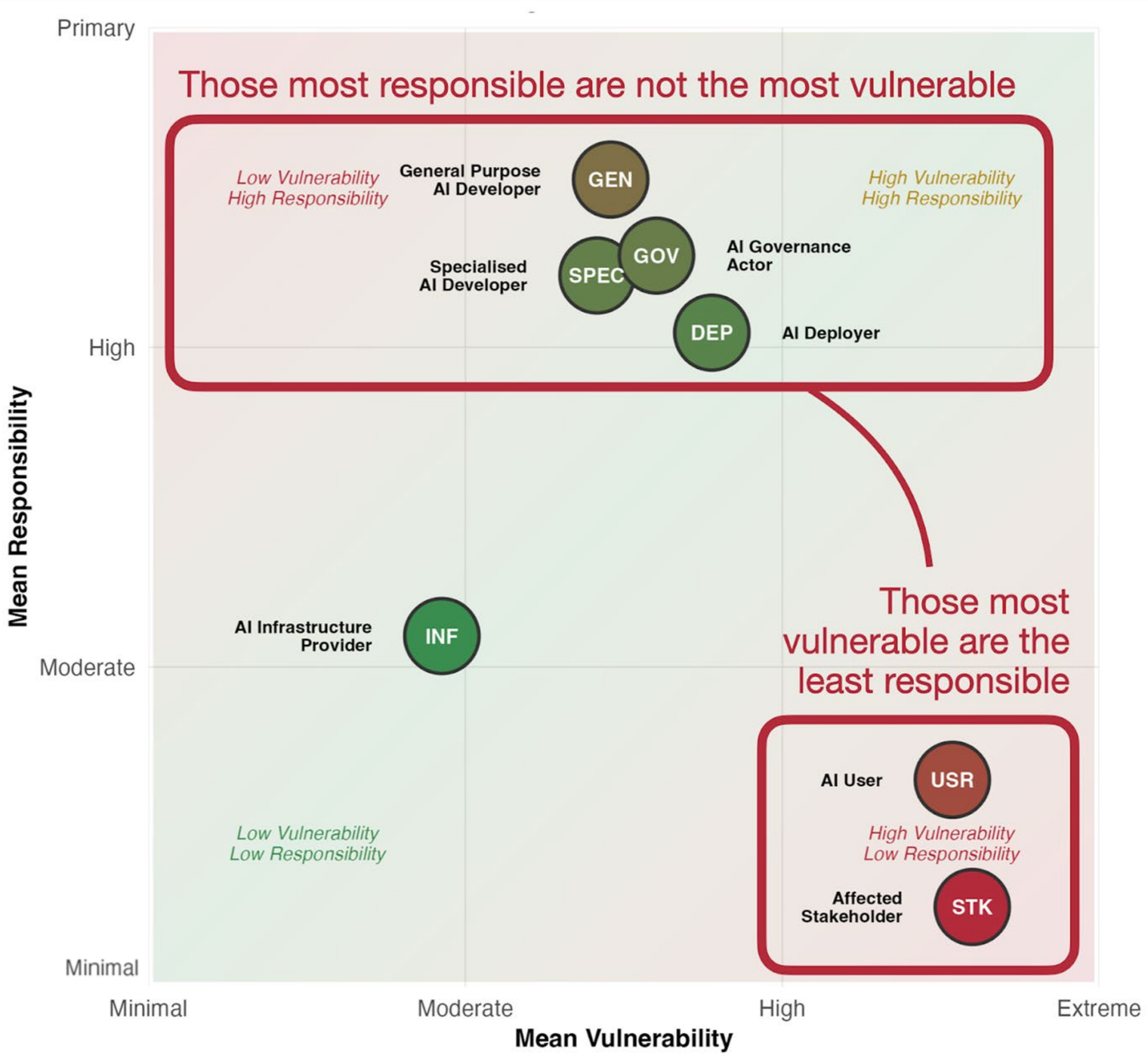
Highly responsible

Moderately responsible

Can those who are vulnerable to AI risks take responsibility for addressing them?



Can those who are vulnerable to AI risks take responsibility for addressing them?



How are global companies responding to the top AI risks?

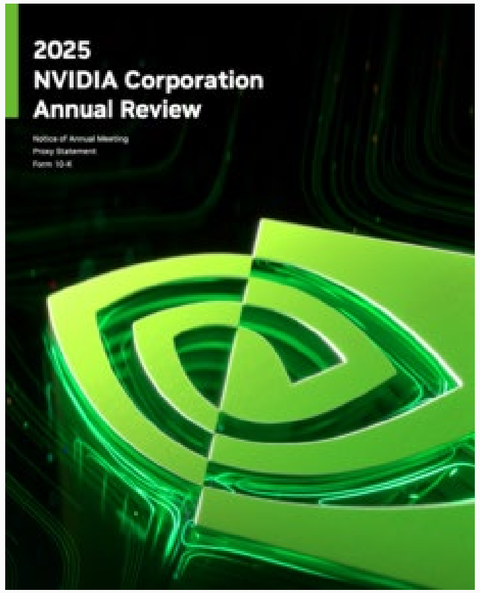
Examples of companies included

+ 150 more

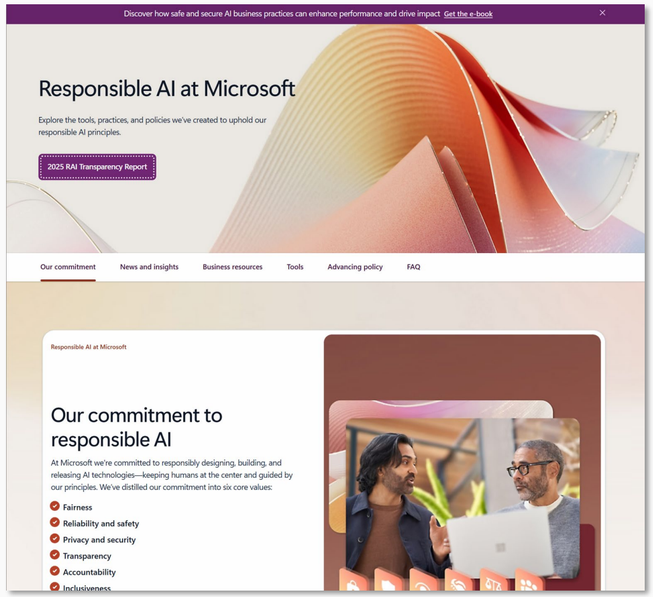
What we found in our search

13,000 public documents on AI

1-16 documents on AI risks & mitigations per company



Annual reports



Webpages

Responsible AI frameworks & policies

How we determined if a company responded to a risk

Protecting against cyber risk

AI and cyber risk are closely connected. Bad actors can exploit AI to scale their operations and increase the effectiveness of attacks, such as scams or phishing, while AI itself can introduce new cyber vulnerabilities…

Discrimination & Toxicity	Discrimination	Toxic content	Unequal performance
Privacy & Security	Loss of privacy	AI security vulnerabilities	
Misinformation	False information	Loss of consensus reality	
Malicious Actors & Misuse	Disinformation & influence	AI weapons & cyberattacks	**AI fraud & scams**
Human-Computer Interaction	Overreliance & unsafe use	Loss of human agency	
Socioeconomic & Environmental Harm	Power Centralization	Inequality & unemployment	Devaluation of human creativity
	Governance failure	Competitive dynamics	Environmental harm
AI System Safety, Failures, & Limitations	AI misalignment	Capability & robustness	AI welfare
	Dangerous capabilities	Transparency & interpretability	Multi-agent risks

What top AI risks is each sector responding to?

Company	Cov	Docs	Discrimination & Toxicity			Privacy & Security		Misinformatio		Malicious actors			Human-Computer Interaction		Socioeconomic & Environmental						AI system safety, failures, & limitations					
			1.1 Discriminat	1.2 Toxic content	1.3 Unequal perfor-mance	2.1 Loss of privacy	2.2 AI security vulnerabili-ties	3.1 False informa-tion	3.2 Loss of consensus reality	4.1 Disinformat & influenc	4.2 AI weap-ons & cyberat-tacks	4.3 AI fraud & scams	5.1 Overreliand & unsafe use	5.2 Loss of human agency	6.1 Power centraliza-tion	6.2 Inequality & unemploy-ment	6.3 Devaluatio of human creativity	6.4 Competitiv dynamics	6.5 Governanc failure	6.6 Environmer harm	7.1 AI misalign-ment	7.2 Dangerous capabili-ties	7.3 Capability & robust-ness	7.4 Transparen & interpretab ity	7.5 AI welfare	7.6 Multi-agent risks
▶Information (42)			55%	21%	7%	40%	24%	29%	—	19%	55%	36%	14%	10%	—	5%	52%	2%	12%	38%	5%	7%	19%	19%	—	2%
▶Finance and Insurance (40)			38%	—	—	18%	8%	15%	—	8%	28%	40%	3%	—	—	5%	10%	—	3%	5%	—	—	13%	18%	—	3%
▶Professional and Technical Services (2)			50%	—	—	50%	—	50%	—	50%	100%	—	—	—	—	—	50%	—	50%	50%	—	—	50%	—	—	—
▶Trade, Transportation, and Utilities (34)			15%	3%	—	6%	6%	3%	—	3%	9%	9%	—	—	—	—	9%	—	—	—	—	3%	6%	9%	—	—
▶Management, Administrative, and Support Services (3)			—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
▶Real Estate and Rental and Leasing (2)			—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
▶Manufacturing - Consumer Products (10)			30%	—	10%	10%	—	20%	—	—	10%	10%	10%	—	—	—	20%	—	10%	—	—	—	10%	10%	—	—
▶Manufacturing - Health and Life Sciences (20)			50%	—	—	30%	—	—	—	5%	10%	10%	—	—	—	—	—	—	5%	10%	—	—	10%	30%	—	—
▶Manufacturing - Industrial Equipment and Defense (17)			12%	—	—	6%	12%	6%	—	—	18%	—	—	—	—	—	6%	—	—	6%	—	—	—	—	—	—
▶Mining and Manufacturing - Energy and Resources (24)			—	—	—	—	4%	—	—	4%	8%	4%	—	—	—	—	—	—	—	—	—	—	—	—	—	—
▶Accommodation, Food, and Other Services (2)			50%	—	—	—	—	50%	—	—	50%	—	—	—	—	—	50%	—	—	—	—	—	—	—	—	—

Preliminary findings: expected release in the next two months.

Extremely vulnerable

Highly vulnerable

% = proportion of companies in sector responding to specific risk
Sectors with <5 companies in sample not displayed

Please help us with data validation

- Validation emails are being sent from this week to included companies
- Please scan this QR code or email airisk@ mit.edu to participate

airisk.mit.edu/preview/orgrev-feedback

Gaps and opportunities

Gaps we would like help to fill

- **What are organizations doing privately?**
  - Extending the organizational review to survey organizations to understand what they're doing internally

- **Which mitigations are the best fit for specific risks?**
  - Consulting with experts to evaluate mitigations for implementation feasibility and cost

- **What are labs and governments doing to reduce the risk to AI?**
  - Extending our organizational review to cover frontier AI developers and governance actors

Gaps NIST might help close

- **Definitions and shared understanding**
  - What should count as an AI incident, hazard, or near-miss?
- **Reporting**
  - How can we move beyond voluntary disclosures and media reports?
  - How can we get better data to understand and predict incident trends?
- **Preparedness**
  - How can we prepare for more serious AI failures, misuse, and agentic systems?
  - How can we help organizations to learn from incidents and update governance as risks evolve?
- **Coordination and responsibility:**
  - Who should act across the AI value chain when an incident occurs?
  - Where do we need information sharing and joint investigations?

Thank you!

Neil Thompson

airisk.mit.edu
airisk@ mit.edu