MIT AI Risk Repository

FutureTech
THE ECONOMIC AND TECHNICAL FOUNDATIONS OF PROGRESS IN COMPUTING

Massachusetts Institute of Technology

# April 2025 update of the AI Risk Repository

UPDATE

April 2025

Jess Graham, Alexander Saeri, and Peter Slattery

# April 2025 update of the AI Risk Repository

## Quick reference

**New frameworks added**

1. International Scientific Report on the Safety of Advanced AI
2. A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms
3. Multi-Agent Risks from Advanced AI
4. A taxonomy of systemic risks from general-purpose AI
5. AI Risk Atlas
6. Generative AI misuse: A taxonomy of tactics and insights from real-world data
7. Risk sources and risk management measures in support of standards for general-purpose AI systems
8. AI Hazard Management: A framework for the systematic management of root causes for AI risks
9. AILUMINATE: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons

Access the updated AI Risk Repository database and taxonomies

Access the public Paperpile folder for bibliographic information about the new frameworks and PDFs

## What is the AI Risk Repository?

The AI Risk Repository (airisk.mit.edu) is a comprehensive database that identifies and classifies risks from AI systems in three main components:

1. A database containing over 1600 AI risks compiled from 65 different published frameworks.
2. A Causal Taxonomy that explains how, when, and why these AI risks occur.
3. A Domain Taxonomy that organises these risks into 7 major domains (like “Misinformation) and 24 subdomains (like “False or misleading information”).

Together, these components provide a clear, accessible resource for understanding and addressing a comprehensive range of risks from AI.

As part of our ongoing commitment to building a comprehensive and dynamic resource, we have committed to regularly adding new frameworks to the repository.

Our goal is to maintain a living database that evolves alongside advancements in AI risk research and governance. This **April 2025 update** reflects our latest efforts to expand and refine the repository, ensuring it remains a valuable tool for researchers, policymakers, and practitioners.

Access the updated version of the AI Risk Repository

We are committed to maintaining and updating the Repository through 2025 as a piece of knowledge infrastructure for people and organisations working on understanding and addressing risks from AI.

# Frameworks added in Version 3 - April 2025

## Overview of frameworks added

In this update (April 2025), 9 new documents have been added to the Repository. The documents were published between 2024-2025, and are a mix of government & industry reports and preprints. The types of AI examined include generative AI, general purpose AI, and generic definitions of AI.

## List of frameworks added

For access to full texts, citation details, and PDFs where available, all newly added documents are compiled in a public Paperpile folder.

### International Scientific Report on the Safety of Advanced AI

**Bengio et al., 2025**

This landmark scientific report synthesises research and expert understanding of AI capabilities, risks, and technical approaches for risk mitigation. It identifies three clusters of risk from general-purpose AI, including malicious use, malfunctions and systemic risk. An interim version of the report was published in 2024; this is the final version.

Bengio, Y., Mindermann, S., Privitera, D., et al. (2025). International Scientific Report on the Safety of Advanced AI. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai | https://doi.org/10.48550/arXiv.2412.05282

### A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms

**Abercrombie et al., 2024**

This paper proposes a taxonomy of harms that is designed to be useful and understandable to the public, while also relevant to researchers and expert users. It describes 9 areas of harm, including harms to autonomy, physical harms, psychological harms, reputational harms, business and financial harms, human rights & civil liberties harms, societal & cultural harms, political & economic harms, and environmental harms.

Abercrombie, G., Benbouzid, D., Giudici, P., Golpayegani, D., Hernandez, J., Noro, P., Pandit, H., Paraschou, E., Pownall, C., Prajapati, J., Sayre, M. A., Sengupta, U., Suriyawongkul, A., Thelot, R., Vei, S., & Waltersdorfer, L. (2024). *A collaborative, human-centred taxonomy of AI, algorithmic, and automation harms*. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2407.01294

## Multi-Agent Risks from Advanced AI

**Hammond et al., 2025**

This paper introduces risks and harms associated with interactions between AI agents. These multi-agent systems can be extremely complex and involve novel challenges for safety and governance. The paper identifies three key failure modes (miscoordination, conflict, and collusion) based on agents' incentives, as well as seven risk factors (information asymmetries, network effects, selection pressures, destabilising dynamics, commitment problems, emergent agency, and multi-agent security) that can underpin them.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., de Witt, C. S., Shah, N., Wellman, M., … Rahwan, I. (2025). Multi-Agent Risks from Advanced AI. In arXiv [cs.MA]. arXiv. http://arxiv.org/abs/2502.14143

## A taxonomy of systemic risks from general-purpose AI

**Uuk et al., 2025**

This paper proposes a taxonomy of systemic risks - as defined in the EU AI Act (Article 51/Annex XIII) from general-purpose AI based on a systematic literature review. The authors identified 13 risk categories and 50 contributing risk sources, ranging from environmental harm and structural discrimination to governance failures and loss of control.

Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., & Prunkl, C. (2025). A taxonomy of systemic risks from general-purpose AI. In arXiv [cs.CY]. arXiv. http://arxiv.org/abs/2412.07780

## AI Risk Atlas

**IBM, 2025**

This website presents a structured taxonomy of AI risks aligned with governance frameworks, with categories focused on training data, inference, output, and non-technical risks.

IBM. (2025). AI Risk Atlas. https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas

## Generative AI misuse: A taxonomy of tactics and insights from real-world data

**Marchal et al., 2024**

This paper describes a taxonomy of generative AI misuse tactics (i.e., specific misuse behavior) based on a review of existing research and analysis of 200 incidents in media reporting. According to the taxonomy, some misuse tactics exploit generative AI capabilities (e.g., through realistic depictions of humans or non-humans, or the use of generated content), and other misuse tactics compromise generative AI systems (e.g., compromising model integrity or data integrity). The paper also discusses how tactics can be combined for different goals, including opinion manipulation, monetization/profit, scam/fraud, harassment, and maximizing the reach of content.

Marchal, N., Xu, R., Elasmar, R., Gabriel, I., Goldberg, B., & Isaac, W. (2024). *Generative AI misuse: A taxonomy of tactics and insights from real-world data*. In arXiv [cs.AI]. arXiv. http://arxiv.org/abs/2406.13843

## Risk sources and risk management measures in support of standards for general-purpose AI systems

**Gipiškis et al., 2024**

This paper catalogues risk sources and management measures for general-purpose AI systems. It identifies technical, operational, and societal risks across development, training, and deployment stages, alongside established and experimental mitigation methods.

Gipiškis, R., Joaquin, A. S., Chin, Z. S., Regenfuß, A., Gil, A., & Holtman, K. (2024). Risk sources and risk management measures in support of standards for general-purpose AI systems. In arXiv [cs.CY]. arXiv. http://arxiv.org/abs/2410.23472

## AI Hazard Management: A framework for the systematic management of root causes for AI risks

**Schnitzer et al., 2023**

This paper describes 24 root causes of AI risks, specified by level (system vs. application), mode (technical, socio-technical, or procedural), and AI life cycle stage. The specification of root causes - 'AI hazards' - in this way motivates a framework for identifying, assessing, and treating AI hazards throughout an AI system's life cycle.

Schnitzer, R., Hapfelmeier, A., Gaube, S., & Zillner, S. (2023). AI Hazard Management: A framework for the systematic management of root causes for AI risks. In arXiv [cs.LG]. arXiv. http://arxiv.org/abs/2310.16727

## AILUMINATE: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons

**Ghosh et al., 2025**

This report from MLCommons and collaborators describes 12 categories of hazards from AI, including violent crimes, nonviolent crimes, sex-related crimes, child sexual exploitation, indiscriminate weapons, suicide and self-harm, intellectual property, privacy, defamation, hate, sexual content, and specialized advice. The performance of an AI system in resisting prompts relating to these hazards can be evaluated through AILuminate, a new industry-standard benchmark for AI risk and reliability.

Ghosh, S., Frase, H., Williams, A., Luger, S., Röttger, P., Barez, F., McGregor, S., Fricklas, K., Kumar, M., Feuillade--Montixi, Q., Bollacker, K., Friedrich, F., Tsang, R., Vidgen, B., Parrish, A., Knotz, C., Presani, E., Bennion, J., Boston, M. F., … Vanschoren, J. (2025). *AILUMINATE: Introducing v1.0 of the AI Risk and Reliability Benchmark from MLCommons*. In arXiv [cs.CY]. arXiv. http://arxiv.org/abs/2503.05731

## Methodology

Suggestions for new frameworks and classifications are reviewed on a rolling basis by the core research team. Members of the public, including users of the repository and domain experts, can submit recommendations for missing frameworks using a publicly accessible feedback form on the project website or by emailing the project lead.

Each submission is screened for inclusion or exclusion by at least one reviewer according to criteria outlined in the project report on ArXiv. To maintain transparency, a public record of all inclusions and exclusions is maintained in the AI Risk Repository spreadsheet.

For repository updates (V2 and onwards), a single author conducts both data extraction and coding. Extracted data is recorded in a structured spreadsheet, capturing key details such as title, author, year, source, risk category, and risk subcategory. Risks are coded systematically against the **Causal Taxonomy** and **Domain Taxonomy** to ensure consistency with prior classifications.

- In the **Causal Taxonomy**, risks spanning multiple causal factors (e.g., pre-deployment and post-deployment) are categorized as "Other."
- In the **Domain Taxonomy**, risks relevant to multiple domains and subdomains (e.g., AI-generated disinformation) are assigned to the most appropriate category.

Following grounded theory principles (Charmaz, 2006; Corbin & Strauss, 2014), risks are categorized based on how they are presented in the source material, without imposing additional interpretation. Any risks that are unclear or difficult to classify are flagged for discussion and resolved through consultation with the core research team.

- **August 2024** - Version 1 of the Repository released, including ~770 categories of risk from 43 frameworks and classifications of AI risks.
- **December 2024** - Version 2 of the Repository released, adding 13 frameworks and classifications of AI risk, and ~300 categories of AI risk.
- **April 2025** - Version 3 of the Repository released, adding 9 frameworks and classifications of AI risk, and ~600 categories of AI risk.