MIT AI Risk Repository

FutureTech
THE ECONOMIC AND TECHNICAL FOUNDATIONS OF PROGRESS IN COMPUTING

Massachusetts Institute of Technology

December 2024 update of the AI Risk Repository

A summary of the 13 new frameworks added to the repository

UPDATES

December, 2024

Authors

Peter Slattery, Alexander K. Saeri and Jess Graham.

What is the AI Risk Repository?

The AI Risk Repository is a comprehensive database that identifies and classifies risks from AI systems in three main components:

1. A database containing over 1000 AI risks compiled from 56 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 23 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 **December 2024 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 here.

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. We intend to share an update each quarter in 2025 with (1) new frameworks added to the repository, and (2) changes in risk definitions based on new frameworks.

A stretch goal for 2025 is a major update to the Repository, which could involve adding or removing categories of risk.

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 preprint. 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.

Overview of Added Frameworks

Following **the December 2024 repository update**, 13 new documents have been added to the Repository. The documents were published between 2018-2024, and are a mix of government & industry reports, peer reviewed journal articles, and preprints, with authors from US, UK, Australia, Canada, China, and Germany. The types of AI examined include generative AI, large language models, and “Artificial General Intelligence”, in addition to generic definitions of AI.

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 (interim report)
Bengio et al., 2024

This 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, as well as cross-cutting factors that exacerbate risks.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Goldfarb, D., Heidari, H., Khalatbari, L., Longpre, S., Mavroudis, V., Mazeika, M., Ng, K. Y., Okolo, C. T., Raji, D., Skeadas, T., & Tramèr, F. (2024). *International Scientific Report on the Safety of Advanced AI*. https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai

AI risk categorization decoded (AIR 2024): From government regulations to corporate policies
Zeng et al., 2024
This preprint systematically reviews US, EU, and China AI regulations, as well as 16 leading AI developers' policies, to construct an AI risk taxonomy with more than 300 categories of risks. The authors compare developers' and countries/jurisdictions' responses to AI risks.

Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., Song, D., Liang, P., & Li, B. (2024). AI risk categorization decoded (AIR 2024): From government regulations to corporate policies. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2406.17864

A survey of the potential long-term impacts of AI
Clarke & Whittlestone, 2022
This preprint reviews research on the societal impacts of AI. Authors describe both positive and negative implications for AI in science, cooperation, power, epistemics, and values.

Clarke, S., & Whittlestone, J. (2022). A survey of the potential long-term impacts of AI. In *arXiv [cs.CY]*. arXiv. https://doi.org/10.1145/3514094.3534131

Harm to nonhuman animals from AI: A systematic account and framework
Coghlan & Parker, 2023

This journal article reviews research on how AI could harm nonhuman animals and describes a framework for organising the harms, including intentional harms (socially accepted, socially condemned); unintentional harms (direct, indirect); and foregone benefits for animals.

Coghlan, S., & Parker, C. (2023). Harm to nonhuman animals from AI: A systematic account and framework. *Philosophy & Technology*, *36*(2), 1–34. https://doi.org/10.1007/s13347-023-00627-6

AGI Safety Literature Review

Everitt et al., 2018

This preprint reviews technical research about the (then-emerging) field of artificial general intelligence (AGI) safety, including problems in designing or building safe AGI, risks of unsafe AGI, and proposed solutions.

Everitt, T., Lea, G., & Hutter, M. (2018). AGI Safety Literature Review. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1805.01109

GenAI against humanity: nefarious applications of generative artificial intelligence and large language models

Ferrara, 2024

This journal article reviews research on the misuse of generative AI and LLMs to harm humans, and presents a matrix that maps three malicious intents (dishonesty, propaganda, deception) against four types of harm (personal, financial, informational, and socio-technical).

Ferrara, E. (2024). GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, *7*(1), 549–569. https://doi.org/10.1007/s42001-024-00250-1

Future Risks of Frontier AI

Government Office for Science (UK), 2023

This report summarises the capabilities and risks of highly capable general-purpose AI systems ("frontier AI"), including through scenario analysis exploring five areas of key uncertainty: capability, ownership & access, safety, level & distribution of use, and geopolitical context.

Government Office for Science (UK). (2023). *Future Risks of Frontier AI*. Government Office for Science. https://assets.publishing.service.gov.uk/media/653bc393d10f3500139a6ac5/future-risks-of-frontier-ai-annex-a.pdf

Regulating under uncertainty: Governance options for generative AI

G'sell, 2024

This report from the Stanford Cyber Policy Center explores governance approaches for generative AI, including self-regulation, co-regulation, and government regulation.

G'sell, F. (2024). Regulating under uncertainty: Governance options for generative AI. In *Social Science Research Network*. https://doi.org/10.2139/ssrn.4918704

Ten hard problems in artificial intelligence we must get right

Leech et al., 2024

This preprint reviews research about and proposes approaches to address 10 "hard problems" (Schmidt Futures, "AI2050") that must be solved to realise the benefits of AI, including capabilities, assurance, alignment, application, economic disruption, inclusion, responsible deployment, geopolitical disruption, governance, and human meaning.

Leech, G., Garfinkel, S., Yagudin, M., Briand, A., & Zhuravlev, A. (2024). Ten hard problems in artificial intelligence we must get right. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2402.04464

Advanced AI governance: A literature review of problems, options, and proposals

Maas, 2023

This report from the Institute for Law & AI provides an structured overview of research in governance of 'advanced AI', covering three key areas: the challenges of governing advanced AI; the options for governing advanced AI, including actors, levers, and pathways of influence; proposed policies to govern advanced AI.

Maas, M. M. (2023). *Advanced AI governance: A literature review of problems, options, and proposals*. Institute for Law & AI. https://doi.org/10.2139/ssrn.4629460

Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks

Maham & Küspert, 2023

This report from Stiftung Neue Verantwortung (now Interface) describes three categories of risk from general-purpose AI: Unreliability, Misuse, and Systemic Risks, each of which includes three specific risks. The risks are illustrated with examples and scenarios, and the report makes recommendations for how EU policymakers could act to respond to the diverse risks from general purpose AI, including and beyond the EU AI Act.

Maham, P., & Küspert, S. (2023). *Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks*. Stiftung Neue Verantwortung. https://www.interface-eu.org/publications/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks

Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)

National Institute of Standards and Technology (USA), 2024

This report is the generative AI profile for the NIST AI Risk Management Framework (AI RMF 1.0). It defines and describes risks that are relevant to generative AI, and proposes actions to address these risks aligned with the NIST RMF (govern, map, measure, and manage).

National Institute of Standards and Technology (US). (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)*. National Institute of Standards and Technology (US). https://doi.org/10.6028/nist.ai.600-1

AI Safety Governance Framework

National Technical Committee 260 on Cybersecurity of SAC (China), 2024

This report is version 1 of China's AI Safety Governance Framework, which describes principles for governing AI safety governance; distinguishes between inherent AI safety risks and safety risks in AI applications; describes options for both technological and governance approaches to address these risks; and sets out safety guidelines for AI development and application.

National Technical Committee 260 on Cybersecurity of SAC. (2024). *AI Safety Governance Framework*. https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf