TechXplore

Global AI adoption is outpacing risk understanding, researchers warn

August 15 2024, by Peter Slattery, Rachel Gordon, and Neil Thompson

Credit: Massachusetts Institute of Technology

As organizations rush to implement artificial intelligence (AI), a new analysis of AI-related risks finds significant gaps in our understanding, highlighting an urgent need for a more comprehensive approach.

The adoption of AI is rapidly increasing; census data shows a significant (47%) rise in AI usage within U.S. industries, jumping from 3.7% to

5.45% between September 2023 and February 2024.

However, a comprehensive review from researchers at MIT CSAIL and MIT FutureTech has uncovered critical gaps in existing AI risk frameworks. Their analysis reveals that even the most thorough individual framework overlooks approximately 30% of the risks identified across all reviewed frameworks.

To help address this, they have collaborated with colleagues from the University of Queensland, Future of Life Institute, KU Leuven, and Harmony Intelligence, to release the first-ever AI Risk Repository: a comprehensive and accessible living database of 700+ risks posed by AI that will be expanded and updated to ensure that it remains current and relevant.

"Since the AI risk literature is scattered across peer-reviewed journals, preprints, and industry reports, and quite varied, I worry that decision-makers may unwittingly consult incomplete overviews, miss important concerns, and develop collective blind spots," says Dr. Peter Slattery, an incoming postdoc at the MIT FutureTech Lab and current project lead.

After searching several academic databases, engaging experts, and retrieving more than 17,000 records, the researchers identified 43 existing AI risk classification frameworks. From these, they extracted more than 700 risks. They then used approaches that they developed from two existing frameworks to categorize each risk by cause (e.g., when or why it occurs), risk domain (e.g., "Misinformation"), and risk subdomain (e.g., "False or misleading information").

Examples of risks identified include "Unfair discrimination and misrepresentation," "Fraud, scams, and targeted manipulation," and "Overreliance and unsafe use." More of the risks analyzed were attributed to AI systems (51%) than humans (34%) and presented as

emerging after AI was deployed (65%) rather than during its development (10%).

The most frequently addressed risk domains included "AI system safety, failures, and limitations" (76% of documents); "Socioeconomic and environmental harms" (73%); "Discrimination and toxicity" (71%); "Privacy and security" (68%); and "Malicious actors and misuse" (68%). In contrast, "Human-computer interaction" (41%) and "Misinformation" (44%) received comparatively less attention.

Some risk subdomains were discussed more frequently than others. For example, "Unfair discrimination and misrepresentation" (63%), "Compromise of privacy" (61%), and "Lack of capability or robustness" (59%), were mentioned in more than 50% of documents. Others, like "AI welfare and rights" (2%), "Pollution of information ecosystem and loss of consensus reality" (12%), and "Competitive dynamics" (12%), were mentioned by less than 15% of documents.

On average, frameworks mentioned just 34% of the 23 risk subdomains identified, with nearly a quarter covering less than 20%. No document or overview mentioned all 23 risk subdomains, and the most comprehensive (Gabriel et al, 2024) covered only 70%.

The work addresses the urgent need to help decision-makers in government, research, and industry understand and prioritize the risks associated with AI and work together to address them. "Many AI governance initiatives are emerging across the world focused on addressing key risks from AI," says collaborator Risto Uuk, EU Research Lead at the Future of Life Institute. "These institutions need a more comprehensive and complete understanding of the risk landscape."

Researchers and risk evaluation professionals are also impeded by the fragmentation of current literature. "It is hard to find specific studies of

risk in some niche domains where AI is used, such as weapons and military decision support systems," explains Taniel Yusef, a Research Affiliate, at The Center for the Study of Existential Risk, at the University of Cambridge who was not involved in the research. "Without referring to these studies, it can be difficult to speak about technical aspects of AI risk to non-technical experts. This repository helps us do that."

"There's a significant need for a comprehensive database of risks from advanced AI which safety evaluators like Harmony Intelligence can use to identify and catch risks systematically," argues collaborator Soroush Pour, CEO & Co-founder of AI safety evaluations and red teaming company Harmony Intelligence. "Otherwise, it's unclear what risks we should be looking for, or what tests need to be done. It becomes much more likely that we miss something by simply not being aware of it."

AI's risky business

The researchers built on two frameworks (Yampolskiy 2016 & Weidinger et al, 2022) in categorizing the risks they extracted. Based on these approaches, they group the risks in two ways.

First by causal factors:

1. Entity: Human, AI, and Other;
2. Intentionality: Intentional, Unintentional, and Other; and
3. Timing: Pre-deployment; Post-deployment, and Other.

Second, by seven AI risk domains:

1. Discrimination & toxicity,
2. Privacy & security,
3. Misinformation,

4. Malicious actors & misuse,
5. Human-computer interaction,
6. Socioeconomic & environmental, and
7. AI system safety, failures, & limitations.

These are further divided into 23 subdomains (full descriptions here):

- 1.1. Unfair discrimination and misrepresentation
- 1.2. Exposure to toxic content
- 1.3. Unequal performance across groups
- 2.1. Compromise of privacy by leaking or correctly inferring sensitive information
- 2.2. AI system security vulnerabilities and attacks
- 3.1. False or misleading information
- 3.2. Pollution of the information ecosystem and loss of consensus reality
- 4.1. Disinformation, surveillance, and influence at scale
- 4.2. Cyberattacks, weapon development or use, and mass harm
- 4.3. Fraud, scams, and targeted manipulation
- 5.1. Overreliance and unsafe use
- 5.2. Loss of human agency and autonomy
- 6.1. Power centralization and unfair distribution of benefits
- 6.2. Increased inequality and decline in employment quality
- 6.3. Economic and cultural devaluation of human effort
- 6.4. Competitive dynamics
- 6.5. Governance failure
- 6.6. Environmental harm
- 7.1. AI pursuing its own goals in conflict with human goals or values
- 7.2. AI possessing dangerous capabilities
- 7.3. Lack of capability or robustness
- 7.4. Lack of transparency or interpretability
- 7.5. AI welfare and rights

"The AI Risk Repository is, to our knowledge, the first attempt to rigorously curate, analyze, and extract AI risk frameworks into a publicly accessible, comprehensive, extensible, and categorized risk database. It is part of a larger effort to understand how we are responding to AI risks and to identify if there are gaps in our current approaches," says Dr. Neil Thompson, head of the MIT FutureTech Lab and one of the lead researchers on the project.

"We are starting with a comprehensive checklist, to help us understand the breadth of potential risks. We plan to use this to identify shortcomings in organizational responses. For instance, if everyone focuses on one type of risk while overlooking others of similar importance, that's something we should notice and address."

The next phase will involve experts evaluating and prioritizing the risks within the repository, then using it to analyze public documents from influential AI developers and large companies. The analysis will examine if organizations respond to risks from AI—and do so in proportion to experts' concerns—and compare risk management approaches across different industries and sectors.

The repository is freely available online to download, copy, and use. Feedback and suggestions can be submitted here.

More information: The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence

Provided by Massachusetts Institute of Technology