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

Peter Slattery[1,2], Alexander K. Saeri[1,2], Emily A. C. Grundy[1,2], Jess Graham[3], Michael Noetel[2,3], Risto Uuk[4,5], James Dao[6], Soroush Pour[6], Stephen Casper[7], and Neil Thompson[1].

[1]MIT FutureTech, Massachusetts Institute of Technology, [2]Ready Research, [3]School of Psychology, The University of Queensland, [4]Future of Life Institute, [5]KU Leuven, [6]Harmony Intelligence, [7]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Correspondence to pslat@mit.edu.

Abstract

The risks posed by Artificial Intelligence (AI) are of considerable concern to academics, auditors, policymakers, AI companies, and the public. However, a lack of shared understanding of AI risks can impede our ability to comprehensively discuss, research, and react to them. This paper addresses this gap by creating an AI Risk Repository to serve as a common frame of reference. This comprises a living database of 777 risks extracted from 43 taxonomies, which can be filtered based on two overarching taxonomies and easily accessed, modified, and updated via our website and online spreadsheets. We construct our Repository with a systematic review of taxonomies and other structured classifications of AI risk followed by an expert consultation. We develop our taxonomies of AI risk using a best-fit framework synthesis. Our high-level Causal Taxonomy of AI Risks classifies each risk by its *causal* factors (1) Entity: Human, AI; (2) Intentionality: Intentional, Unintentional; and (3) Timing: Pre-deployment; Post-deployment. Our mid-level Domain Taxonomy of AI Risks classifies risks into 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. 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. This creates a foundation for a more coordinated, coherent, and complete approach to defining, auditing, and managing the risks posed by AI systems.

Guide for readers

This is a long document. Here are several ways to use this document and its associated materials, depending on your time and interests.

Two-minute engagement

Skim the Plain Language Summary (p. 3).

Ten-minute engagement

Read the Plain Language Summary (p. 3).

Read Insights into the AI Risk Landscape (p. 56), and Implications for key audiences (p. 57).

Policymakers, Model Evaluators & Auditors

Read the Plain Language Summary (p. 3). Skim Detailed descriptions of domains of AI risks (p. 33).

Read Insights into the AI Risk Landscape (p. 56) and the Policymakers and/or Auditors subsections of Implications for key audiences (p. 57).

Researchers

Read the Plain Language Summary (p. 3). Read Figure 1 (p. 15) to understand the methods we used to identify relevant documents and develop two new taxonomies of AI risk; for more detail on how we developed the taxonomies see Best-fit framework synthesis approach (p. 19).

Read Insights into the AI Risk Landscape (p. 56), and the Academics subsection of Implications for key audiences (p. 59) and skim Limitations and directions for future research (p. 60).

Plain Language Summary

- The risks posed by Artificial Intelligence (AI) concern many stakeholders
- Many researchers have attempted to classify these risks
- Existing classifications are uncoordinated and inconsistent
- We review and synthesize prior classifications to produce an AI Risk Repository, including a paper, causal taxonomy, domain taxonomy, database, and website
- To our knowledge, this is the first attempt to rigorously curate, analyze, and extract AI risk frameworks into a publicly accessible, comprehensive, extensible, and categorized risk database

The risks posed by Artificial Intelligence (AI) are of considerable concern to a wide range of stakeholders including policymakers, experts, AI companies, and the public. These risks span various domains and can manifest in different ways: The AI Incident Database now includes over 3,000 real-world instances where AI systems have caused or nearly caused harm.

To create a clearer overview of this complex set of risks, many researchers have tried to identify and group them. In theory, these efforts should help to simplify complexity, identify patterns, highlight gaps, and facilitate effective communication and risk prevention. In practice, these efforts have often been uncoordinated and varied in their scope and focus, leading to numerous conflicting classification systems. Even when different classification systems use similar terms for risks (e.g., “privacy”) or focus on similar domains (e.g., “existential risks”), they can refer to concepts inconsistently. As a result, it is still hard to understand the full scope of AI risk.

In this work, we build on previous efforts to classify AI risks by combining their diverse perspectives into a comprehensive, unified classification system. During this synthesis process, we realized that our results contained two types of classification systems:

- High-level categorizations of causes of AI risks (e.g., when or why risks from AI occur)
- Mid-level hazards or harms from AI (e.g, AI is trained on limited data or used to make weapons)

Because these classification systems were so different, it was hard to unify them; high-level risk categories such as “Diffusion of responsibility” or “Humans create dangerous AI by mistake” do not map to narrower categories like “Misuse” or “Noisy Training Data,” or vice versa. We therefore decided to create two different classification systems that together would form our unified classification system.

The paper we produced and its associated products (i.e., causal taxonomy, domain taxonomy, living database and website) provide a clear, accessible resource for understanding and addressing a comprehensive range of risks associated with AI. We refer to these products as the AI Risk Repository.

What we did

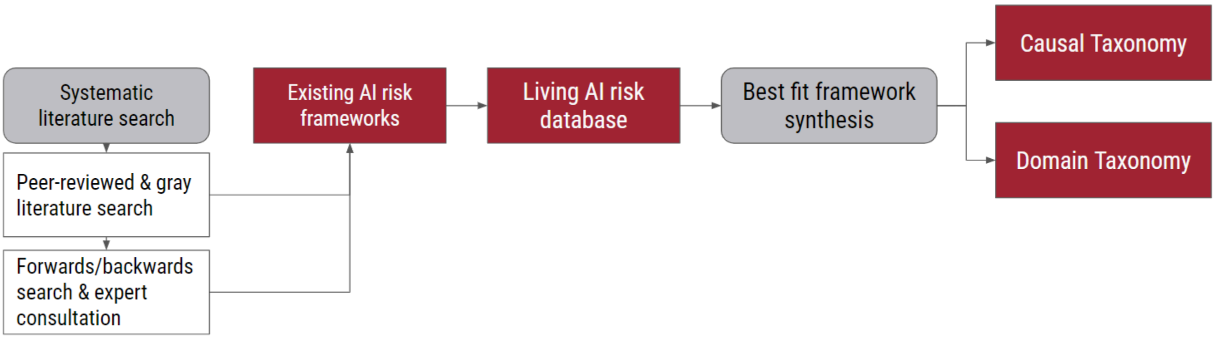
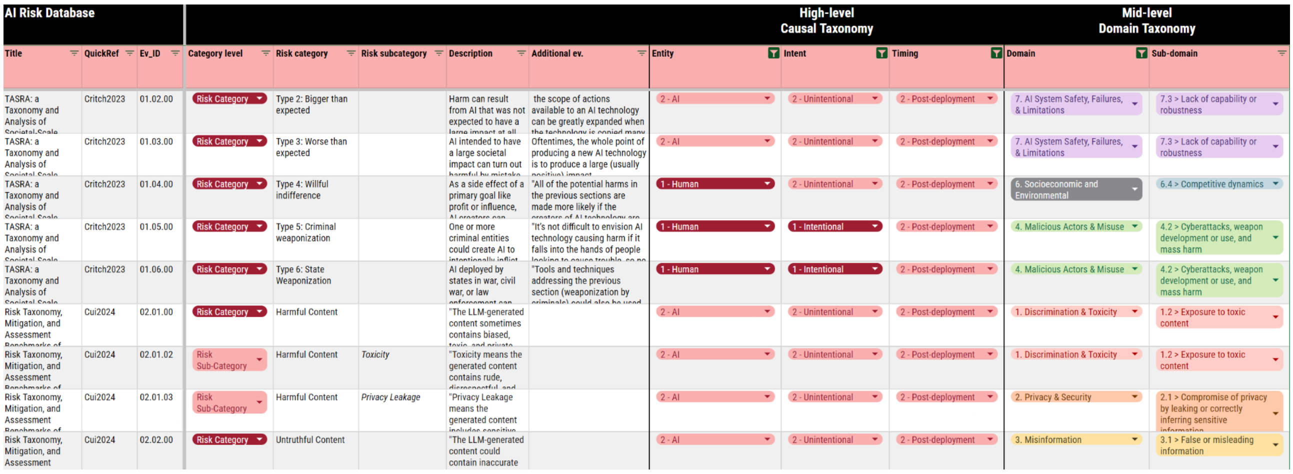

Systematic literature search
Peer-reviewed & gray literature search
Forwards/backwards search & expert consultation
Existing AI risk frameworks
Living AI risk database
Best fit framework synthesis
Causal Taxonomy
Domain Taxonomy

Figure A. Overview of Study Methodology

As shown in Figure A, we used a systematic search strategy, forwards and backwards searching, and expert consultation to identify AI risk classifications, frameworks, and taxonomies. Specifically, we searched several academic databases for relevant research and then used pre-specified rules to define which research would be included in our summary. Next, we consulted experts (i.e., the authors of the included documents) to suggest additional research we should include. Finally, we reviewed i) the bibliographies of the research identified in the first and second stages, and ii) papers that referenced that research to find further relevant research.

At the conclusion of this process, we extracted information about 777 different risks from 43 documents, with quotes and page numbers, into a "living" database we intend to update over time (see Figure B). You can watch an explainer video for the database here.



AI Risk Database								High-level Causal Taxonomy			Mid-level Domain Taxonomy	
Title	QuickRef	Ev_ID	Category level	Risk category	Risk subcategory	Description	Additional ev.	Entity	Intent	Timing	Domain	Sub-domain
TASRA: a Taxonomy and Analysis of Societal-Scale	Critch2023	01.02.00	Risk Category	Type 2: Bigger than expected		Harm can result from AI that was not expected to have a large impact at all	the scope of actions available to an AI technology can be greatly expanded when the technology is copied many	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures, & Limitations	7.3 > Lack of capability or robustness
TASRA: a Taxonomy and Analysis of Societal-Scale	Critch2023	01.03.00	Risk Category	Type 3: Worse than expected		AI intended to have a large societal impact can turn out harmful by mistake	Oftentimes, the whole point of producing a new AI technology is to produce a large (usually positive) impact	2 - AI	2 - Unintentional	2 - Post-deployment	7. AI System Safety, Failures, & Limitations	7.3 > Lack of capability or robustness
TASRA: a Taxonomy and Analysis of Societal-Scale	Critch2023	01.04.00	Risk Category	Type 4: Willful indifference		As a side effect of a primary goal like profit or influence, AI creators can	"All of the potential harms in the previous sections are made more likely if the creators of AI technology are	1 - Human	2 - Unintentional	2 - Post-deployment	6. Socioeconomic and Environmental	6.4 > Competitive dynamics
TASRA: a Taxonomy and Analysis of Societal-Scale	Critch2023	01.05.00	Risk Category	Type 5: Criminal weaponization		One or more criminal entities could create AI to intentionally inflict	"It's not difficult to envision AI technology causing harm if it falls into the hands of people looking to cause trouble, so no	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
TASRA: a Taxonomy and Analysis of Societal-Scale	Critch2023	01.06.00	Risk Category	Type 6: State Weaponization		AI deployed by states in war, civil war, or law enforcement can	"Tools and techniques addressing the previous section (weaponization by criminals) could also be used	1 - Human	1 - Intentional	2 - Post-deployment	4. Malicious Actors & Misuse	4.2 > Cyberattacks, weapon development or use, and mass harm
Risk Taxonomy, Mitigation, and Assessment Benchmarks of	Cui2024	02.01.00	Risk Category	Harmful Content		"The LLM-generated content sometimes contains biased, toxic, and private		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy, Mitigation, and Assessment Benchmarks of	Cui2024	02.01.02	Risk Sub-Category	Harmful Content	*Toxicity*	"Toxicity means the generated content contains rude, disrespectful, and		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy, Mitigation, and Assessment Benchmarks of	Cui2024	02.01.03	Risk Sub-Category	Harmful Content	*Privacy Leakage*	"Privacy Leakage means the generated content includes sensitive		2 - AI	2 - Unintentional	2 - Post-deployment	2. Privacy & Security	2.1 > Compromise of privacy by leaking or correctly inferring sensitive information
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.02.00	Risk Category	Untruthful Content		"The LLM-generated content could contain inaccurate		2 - AI	2 - Unintentional	2 - Post-deployment	3. Misinformation	3.1 > False or misleading information

Figure B. Image of AI Risk Database.

We used a "best fit framework synthesis" approach to develop two taxonomies for classifying these risks. This involved choosing the "best fitting" classification system for our purposes from the set of 43 existing systems we had identified during our search and using this system to categorize the AI risks in our database. Where risks could not be categorized using this system, we updated the existing categories, created new categories, or changed the structure of this system. We repeated this process until we achieved a final version that could effectively code risks in the database.

During coding, we used grounded theory methods to analyze the data. We therefore identified and coded risks as presented in the original sources, without interpretation. Based on this, our Causal Taxonomy groups risks by the entity, intent, and timing presented (see Table A).

Table A. Causal Taxonomy of AI Risks

Category	Level	Description
Entity	Human	The risk is caused by a decision or action made by humans
	AI	The risk is caused by a decision or action made by an AI system
	Other	The risk is caused by some other reason or is ambiguous
Intent	Intentional	The risk occurs due to an expected outcome from pursuing a goal
	Unintentional	The risk occurs due to an unexpected outcome from pursuing a goal
	Other	The risk is presented as occurring without clearly specifying the intentionality
Timing	Pre-deployment	The risk occurs before the AI is deployed
	Post-deployment	The risk occurs after the AI model has been trained and deployed
	Other	The risk is presented without a clearly specified time of occurrence

Our Domain Taxonomy groups risks into seven domains such as discrimination, privacy, and misinformation. These domains are further grouped into 23 risk subdomains (see Table B).

Table B. Domain Taxonomy of AI Risks

Domain / Subdomain	Description
1 *Discrimination & toxicity*	
1.1 Unfair discrimination and misrepresentation	Unequal treatment of individuals or groups by AI, often based on race, gender, or other sensitive characteristics, resulting in unfair outcomes and representation of those groups.
1.2 Exposure to toxic content	AI that exposes users to harmful, abusive, unsafe, or inappropriate content. May involve providing advice or encouraging action. Examples of toxic content include hate speech, violence, extremism, illegal acts, or child sexual abuse material, as well as content that violates community norms such as profanity, inflammatory political speech, or pornography.
1.3 Unequal performance across groups	Accuracy and effectiveness of AI decisions and actions are dependent on group membership, where decisions in AI system design and biased training data lead to unequal outcomes, reduced benefits, increased effort, and alienation of users.
2 *Privacy & security*	
2.1 Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information	AI systems that memorize and leak sensitive personal data or infer private information about individuals without their consent. Unexpected or unauthorized sharing of data and information can compromise user expectation of privacy, assist identity theft, or cause loss of confidential intellectual property.
2.2 AI system security vulnerabilities and attacks	Vulnerabilities that can be exploited in AI systems, software development toolchains, and hardware that results in unauthorized access, data and privacy breaches, or system manipulation causing unsafe outputs or behavior.
3 *Misinformation*	
3.1 False or misleading information	AI systems that inadvertently generate or spread incorrect or deceptive information, which can lead to inaccurate beliefs in users and undermine their autonomy. Humans that make decisions based on false beliefs can experience physical, emotional, or material harms
3.2 Pollution of information ecosystem and loss of consensus reality	Highly personalized AI-generated misinformation that creates “filter bubbles” where individuals only see what matches their existing beliefs, undermining shared reality and weakening social cohesion and political processes.
4 *Malicious actors & misuse*	
4.1 Disinformation, surveillance, and influence at scale	Using AI systems to conduct large-scale disinformation campaigns, malicious surveillance, or targeted and sophisticated automated censorship and propaganda, with the aim of manipulating political processes, public opinion, and behavior.
4.2 Cyberattacks, weapon development or use, and mass harm	Using AI systems to develop cyber weapons (e.g., by coding cheaper, more effective malware), develop new or enhance existing weapons (e.g., Lethal Autonomous Weapons or chemical, biological, radiological, nuclear, and high-yield explosives), or use weapons to cause mass harm.
4.3 Fraud, scams, and targeted manipulation	Using AI systems to gain a personal advantage over others through cheating, fraud, scams, blackmail, or targeted manipulation of beliefs or behavior. Examples include AI-facilitated plagiarism for research or education, impersonating a trusted or fake individual for illegitimate financial benefit, or creating humiliating or sexual imagery.
5 *Human-computer interaction*	
5.1 Overreliance and unsafe use	Anthropomorphizing, trusting, or relying on AI systems by users, leading to emotional or material dependence and to inappropriate relationships with or expectations of AI systems. Trust can be exploited by malicious actors (e.g., to harvest information or enable manipulation), or result in harm from inappropriate use of AI in critical situations (such as a medical emergency). Overreliance on AI systems can compromise autonomy and weaken social ties.

Domain / Subdomain	Description
5.2 Loss of human agency and autonomy	Delegating by humans of key decisions to AI systems, or AI systems that make decisions that diminish human control and autonomy. Both can potentially lead to humans feeling disempowered, losing the ability to shape a fulfilling life trajectory, or becoming cognitively enfeebled.
6 Socioeconomic & environmental harms	
6.1 Power centralization and unfair distribution of benefits	AI-driven concentration of power and resources within certain entities or groups, especially those with access to or ownership of powerful AI systems, leading to inequitable distribution of benefits and increased societal inequality.
6.2 Increased inequality and decline in employment quality	Social and economic inequalities caused by widespread use of AI, such as by automating jobs, reducing the quality of employment, or producing exploitative dependencies between workers and their employers.
6.3 Economic and cultural devaluation of human effort	AI systems capable of creating economic or cultural value through reproduction of human innovation or creativity (e.g., art, music, writing, coding, invention), destabilizing economic and social systems that rely on human effort. The ubiquity of AI-generated content may lead to reduced appreciation for human skills, disruption of creative and knowledge-based industries, and homogenization of cultural experiences.
6.4 Competitive dynamics	Competition by AI developers or state-like actors in an AI “race” by rapidly developing, deploying, and applying AI systems to maximize strategic or economic advantage, increasing the risk they release unsafe and error-prone systems.
6.5 Governance failure	Inadequate regulatory frameworks and oversight mechanisms that fail to keep pace with AI development, leading to ineffective governance and the inability to manage AI risks appropriately.
6.6 Environmental harm	The development and operation of AI systems that cause environmental harm through energy consumption of data centers or the materials and carbon footprints associated with AI hardware.
7 AI system safety, failures & limitations	
7.1 AI pursuing its own goals in conflict with human goals or values	AI systems that act in conflict with ethical standards or human goals or values, especially the goals of designers or users. These misaligned behaviors may be introduced by humans during design and development, such as through reward hacking and goal misgeneralisation, and may result in AI using dangerous capabilities such as manipulation, deception, or situational awareness to seek power, self-proliferate, or achieve other goals.
7.2 AI possessing dangerous capabilities	AI systems that develop, access, or are provided with capabilities that increase their potential to cause mass harm through deception, weapons development and acquisition, persuasion and manipulation, political strategy, cyber-offense, AI development, situational awareness, and self-proliferation. These capabilities may cause mass harm due to malicious human actors, misaligned AI systems, or failure in the AI system.
7.3 Lack of capability or robustness	AI systems that fail to perform reliably or effectively under varying conditions, exposing them to errors and failures that can have significant consequences, especially in critical applications or areas that require moral reasoning.
7.4 Lack of transparency or interpretability	Challenges in understanding or explaining the decision-making processes of AI systems, which can lead to mistrust, difficulty in enforcing compliance standards or holding relevant actors accountable for harms, and the inability to identify and correct errors.
7.5 AI welfare and rights	Ethical considerations regarding the treatment of potentially sentient AI entities, including discussions around their potential rights and welfare, particularly as AI systems become more advanced and autonomous.

What we found

As shown in Table C, most of the risks (51%) were presented as caused by AI systems rather than humans (34%), and as emerging after the AI model has been trained and deployed (65%) rather than before (10%). A similar proportion of risks were presented as intentional (35%) and unintentional (37%)

Table C. AI Risk Database Coded With Causal Taxonomy: Entity, Intent, Timing

Category	Level	Proportion
Entity	Human	34%
	AI	51%
	Other	15%
Intent	Intentional	35%
	Unintentional	37%
	Other	27%
Timing	Pre-deployment	10%
	Post-deployment	65%
	Other	24%

Note. Totals may not match due to rounding.

As shown in Table D, the risk domains that were covered the most in previous documents were:

- AI system safety, failures & limitations - covered in 76% of documents.
- Socioeconomic & environmental harms - covered in 73% of documents.
- Discrimination & toxicity - covered in 71% of documents.

Human-computer interaction (41%) and Misinformation (44%) were less frequently discussed.

No document discussed risks from all 23 subdomains; the highest coverage was 16 out of 23 subdomains (70%; Gabriel et al., 2024). On average, documents mentioned 7 out of 23 (34%) of the AI risk subdomains, with a range of 2 to 16 subdomains mentioned. See Table 9 in the body of the paper for a full breakdown of subdomain coverage by paper.

Some risk subdomains were discussed much more frequently than others, such as:

- Unfair discrimination and misrepresentation (8% of risks).
- AI pursuing its own goals in conflict with human goals or values (8% of risks).
- Lack of capability or robustness (9% of risks).

Some risk subdomains are relatively underexplored, such as:

- AI welfare and rights (<1% of risks).
- Pollution of the information ecosystem and loss of consensus reality (1% of risks).
- Competitive dynamics (1% of risks).

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Stevie Bergman, A., Shelby, R., Marchal, N., Griffin, C., … Manyika, J. (2024). The Ethics of Advanced AI Assistants. In arXiv. Google DeepMind. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/ethics-of-advanced-aiassistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

Table D. AI Risk Database Coded With Domain Taxonomy

Domain / Subdomain	Percentage of risks	Percentage of documents
1 Discrimination & toxicity	***16%***	***71%***
1.1 Unfair discrimination and misrepresentation	8%	63%
1.2 Exposure to toxic content	6%	34%
1.3 Unequal performance across groups	2%	20%
2 Privacy & security	***14%***	***68%***
2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	7%	61%
2.2 AI system security vulnerabilities and attacks	7%	32%
3 Misinformation	***7%***	***44%***
3.1 False or misleading information	5%	39%
3.2 Pollution of information ecosystem and loss of consensus reality	1%	12%
4 Malicious actors & misuse	***14%***	***68%***
4.1 Disinformation, surveillance, and influence at scale	5%	41%
4.2 Cyberattacks, weapon development or use, and mass harm	5%	54%
4.3 Fraud, scams, and targeted manipulation	4%	34%
5 Human-computer interaction	***8%***	***41%***
5.1 Overreliance and unsafe use	5%	24%
5.2 Loss of human agency and autonomy	4%	27%
6 Socioeconomic & environmental harms	***18%***	***73%***
6.1 Power centralization and unfair distribution of benefits	4%	37%
6.2 Increased inequality and decline in employment quality	4%	34%
6.3 Economic and cultural devaluation of human effort	3%	32%
6.4 Competitive dynamics	1%	12%
6.5 Governance failure	4%	32%
6.6 Environmental harm	2%	32%
7 AI system safety, failures & limitations	***24%***	***76%***
7.1 AI pursuing its own goals in conflict with human goals or values	8%	46%
7.2 AI possessing dangerous capabilities	4%	20%
7.3 Lack of capability or robustness	9%	59%
7.4 Lack of transparency or interpretability	3%	27%
7.5 AI welfare and rights	<1%	2%

Note. Domain totals may not match subdomain sums due to rounding and domain-level coding of some risks.

How to use the AI Risk Repository

Our **Database** is free to copy and use. The **Causal and Domain Taxonomies** can be used *separately* to filter this database to identify specific risks, for instance, those focused on risks occurring *pre-deployment* or *post-deployment* or related to a specific risk domain such as *Misinformation*.

The **Causal and Domain Taxonomies** can be used *together* to understand how each causal factor (i.e., *entity*, *intent,* and *timing*) relates to each risk domain or subdomain. For example, a user could filter for *Discrimination & toxicity* risks and use the causal filter to identify the *intentional* and *unintentional* variations of this risk from different sources. Similarly, they could differentiate between sources which examine *Discrimination & toxicity* risks where AI is trained on toxic content *pre-deployment*, and those which examine where AI inadvertently causes harm *post-deployment* by showing toxic content.

We discuss some additional use cases below; see the full paper for more detail.

- General:
  - Onboarding new people to the field of AI risks.
  - A foundation to build on for complex projects.
  - Informing the development of narrower or more specific taxonomies. (e.g., systemic risks, or EU-related misinformation risks).
  - Using the taxonomy for prioritization (e.g., with expert ratings), synthesis (e.g, for a review) or comparison (e.g., exploring public concern across domains).
  - Identifying underrepresented areas (e.g., AI welfare and rights).
- Specific:
  - Policymakers: Regulation and shared standard development.
  - Auditors: Developing AI system audits and standards.
  - Academics: Identifying research gaps and develop education and training.
  - Industry: Internally evaluating and preparing for risks, and developing related strategy, education and training.

How to engage

- Access the Repository via our website: airisk.mit.edu
- Use this form to offer feedback, suggest missing resources or risks, or make contact.

Table 1. 20 most cited documents that present a taxonomy or classification of AI risks

Title	First author	First author affiliation (country)	Year	Type	Citations^	Citations^ / year
Ethical and social risks of harm from language models	Weidinger	Google DeepMind (UK)	2021	Preprint	644	161
Taxonomy of Risks posed by Language Models	Weidinger	Google DeepMind (UK)	2022	Conference Paper	340	113
Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration	Nah	City University of Hong Kong (China)	2023	Journal Article	217	109
The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration	Wirtz	Speyer University (Germany)	2020	Journal Article	209	42
Taxonomy of Pathways to Dangerous Artificial Intelligence	Yampolskiy	University of Louisville (USA)	2016	Journal Article	122	14
Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment	Liu	ByteDance Research (China)	2024	Preprint	102	102
The ethics of ChatGPT -- Exploring the ethical issues of an emerging technology	Stahl	University of Nottingham (UK)	2024	Journal Article	94	94
The risks associated with Artificial General Intelligence: A systematic review	McLean	University Of The Sunshine Coast (Australia)	2023	Journal Article	87	44
Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction	Shelby	Google Research (USA)	2023	Conference Paper	79	40
Model Evaluation for Extreme Risks	Shevlane	Google DeepMind (UK)	2023	Preprint	75	38
An Overview of Catastrophic AI Risks	Hendrycks	Center for AI Safety (USA)	2023	Preprint	71	36
AI Alignment: A Comprehensive Survey	Ji	Peking University (China)	2023	Preprint	66	33
X-Risk Analysis for AI Research	Hendrycks	UC Berkeley (USA)	2022	Preprint	57	19
A Survey of Artificial Intelligence Challenges: Analyzing the Definitions, Relationships, and Evolutions	Saghiri	Tehran Polytechnic (Iran)	2022	Journal Article	56	19
Safety Assessment of Chinese Large Language Models	Sun	Tsinghua University (China)	2023	Preprint	55	28
Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions	Habbal	Karabuk University (Turkiye)	2024	Journal Article	55	55
Governance of artificial intelligence: A risk and guideline-based integrative framework	Wirtz	Speyer University (Germany)	2022	Journal Article	53	18
Evaluating the Social Impact of Generative AI Systems in Systems and Society	Solaiman	Hugging Face (USA)	2023	Preprint	46	23
Towards risk-aware artificial intelligence and machine learning systems: An overview	Zhang	Hong Kong Polytechnic University (China)	2022	Journal Article	43	14
Sociotechnical Safety Evaluation of Generative AI Systems	Weidinger	Google DeepMind (UK)	2023	Preprint	40	20
Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review	Meek	Portland State University (USA)	2016	Conference Paper	38	4

Note. ^ collected from Google Scholar on 28 May 2024. Three organizational/industry reports (AI Verify Foundation, 2023; Allianz Global Corporate & Security, 2018; Electronic Privacy Information Centre, 2023) were not indexed on Google Scholar and are therefore not listed.