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

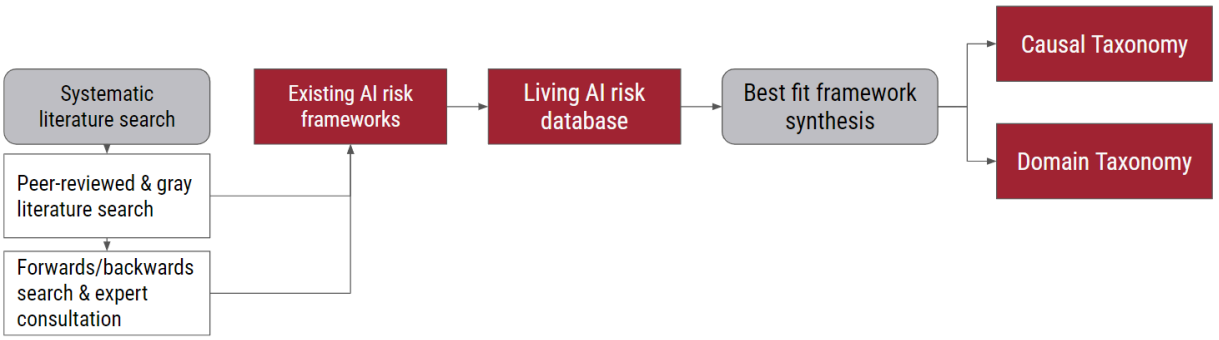


Figure A. Overview of Study Methodology

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	Critch2023	01.02.00	Risk Category	Type 2: Bigger than expected		Harm can result from AI that was not expected to have a	the scope of actions available to an AI technology can be greatly expanded when	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	Critch2023	01.03.00	Risk Category	Type 3: Worse than expected		AI intended to have a large societal impact can turn out	Oftentimes, the whole point of producing a new AI technology is to produce a large (usually	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	Critch2023	01.04.00	Risk Category	Type 4: Willful indifference		As a side effect of a primary goal like profit or influence,	"All of the potential harms in the previous sections are made more likely if the	1 - Human	2 - Unintentional	2 - Post-deployment	6. Socioeconomic and Environmental	6.4 > Competitive dynamics
TASRA: a Taxonomy and Analysis of	Critch2023	01.05.00	Risk Category	Type 5: Criminal weaponization		One or more criminal entities could create AI to	"It's not difficult to envision AI technology causing harm if it falls into the hands of people	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	Critch2023	01.06.00	Risk Category	Type 6: State Weaponization		AI deployed by states in war, civil war, or law	"Tools and techniques addressing the previous section (weaponization by	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	Cui2024	02.01.00	Risk Category	Harmful Content		"The LLM-generated content sometimes contains biased,		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.01.02	Risk Sub-Category	Harmful Content	*Toxicity*	"Toxicity means the generated content contains rude,		2 - AI	2 - Unintentional	2 - Post-deployment	1. Discrimination & Toxicity	1.2 > Exposure to toxic content
Risk Taxonomy, Mitigation, and Assessment	Cui2024	02.01.03	Risk Sub-Category	Harmful Content	*Privacy Leakage*	"Privacy Leakage means the generated content		2 - AI	2 - Unintentional	2 - Post-deployment	2. Privacy & Security	2.1 > Compromise of privacy by leaking or correctly inferring sensitive
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).

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 of Contents

Abstract **1**
Guide for readers **2**
Plain Language Summary **3**
What we did 4
What we found 8
How to use the AI Risk Repository 9
How to engage 10
Table of Contents **11**
Table of Figures and Tables **13**
Introduction **14**
Methods **15**
Systematic literature search 16
Search strategy 16
Extraction into living AI Risk Database 19
Best fit framework synthesis approach 19
Why we developed two taxonomies of AI risk 20
Development of high-level Causal Taxonomy of AI Risks 21
Development of mid-level Domain Taxonomy of AI Risks 22
Coding 23
Results **23**
Systematic literature search 23
Characteristics of included documents 24
Causal Taxonomy of AI Risks 27
Most common causal factors for AI risk 28
Causal factors of AI risk examined by included documents 29
Domain Taxonomy of AI Risks 30
Detailed descriptions of domains of AI risks 33
Most common domains of AI risks 47
Domains of AI risks examined by included documents 48
Subdomains of AI risks examined by included documents 49
Combining the Causal and Domain Taxonomies 52
Most common causal factors for each domain of AI risks 52
Entity x Intent causal factors by each domain of AI risk 54
Discussion **55**
Insights into the “AI risk landscape” 56
Insights from the AI Risk Database and included documents 56
Insights from the Causal Taxonomy 56
Insights from the Domain Taxonomy 57
Implications for key audiences 57
Policymakers 58
Auditors 58
Academics 59
Industry 59

Limitations and directions for future research 60
Review and coding 60
Database and taxonomy 60
Other opportunities for future research 61
Conclusion 61
Appendices 63
Appendix A: Iterative development of Causal Taxonomy and Domain Taxonomy 63
Iterations to develop Causal Taxonomy of AI Risks 63
Iterations to develop Domain Taxonomy of AI risks 66
Appendix B: Characteristics of included documents 69
Declarations 73
References 73

Table of Figures and Tables

Figure A. Overview of Study Methodology 4
Figure B. Image of AI Risk Database. 4
Table A. Causal Taxonomy of AI Risks 5
Table B. Domain Taxonomy of AI Risks 6
Table C. AI Risk Database Coded With Causal Taxonomy: Entity, Intent, Timing 8
Table D. AI Risk Database Coded With Domain Taxonomy 9
Figure 1. Overview of Study Methodology 15
Figure 2. Methodology for Best Fit Framework Synthesis 20
Figure 3. PRISMA Flow Diagram for Systematic Literature Search and Screening 24
Table 1. 20 most cited documents that present a taxonomy or classification of AI risks 26
Table 2. Causal Taxonomy of AI Risks 27
Table 3. AI risk database coded with causal taxonomy: entity, intent, timing 28
Table 4. AI risk database coded with Causal Taxonomy: entity x intent x timing 28
Table 5. Included documents coded with causal taxonomy 29
Table 6. Domain Taxonomy of AI Risks 31
Table 7. AI risk database coded with Domain Taxonomy 47
Table 8. Included documents coded with Domain Taxonomy 48
Table 9. Included documents coded with subdomain taxonomy 50
Table 10. AI Risk Database Coded With Causal Taxonomy and Domain Taxonomy 52
Table 11. AI Risk Database Coded With Causal Taxonomy and Domain Taxonomy: Entity X Intent 54
Table 12. Insights from Causal Taxonomy of AI Risks 57
Table 13. Insights from Domain Taxonomy of AI Risks 57

Introduction

Risks from Artificial Intelligence (AI) are of considerable concern to academics, regulators, policymakers, and the public (Center for AI Safety, 2023; UK Department for Science, Innovation and Technology, 2023a, 2023b). The Responsible AI Collaborative's AI Incident Database now includes over 3,000 real-world instances where AI systems have caused or nearly caused harm (McGregor, 2020). Research and investment in the development and deployment of increasingly capable AI systems has accelerated (Maslej et al., 2024). Concurrent with this attention, researchers and practitioners have sought to understand, evaluate, and address the risks associated with these systems. This work has so far produced a diverse and disparate set of taxonomies, classifications, and other lists of AI risks.

Several examples demonstrate how confusion about AI risks may already impede our effectiveness at risk mitigation. Because there is no canonical set of risks, organizations developing AI are more likely to present risk-mitigation plans that fail to address a comprehensive set of risks (cf. Anthropic, 2023; Google DeepMind, 2024) or lack detail (Anderson-Samways et al., 2024). Similarly, AI-risk evaluators and security professionals are less able to comprehensively evaluate and report on AI risks without a clear understanding of the full range of threats (cf. Nevo et al., 2024). Finally, policymakers may require more extensive support from outside sources (e.g., Department for Science & Technology, 2024) to understand what they need to regulate and legislate.

Another source of confusion is that risks that share a similar name, e.g., "privacy," can refer to different harms or categories of harm (e.g., Meek et al., 2016; Wirtz et al., 2020). Taxonomies that focus on similar domains of risk, e.g., "existential," can vary considerably in their content and how they are constructed (Critch & Russell, 2023; Hendrycks et al., 2023; McLean et al., 2023). Most papers do not base their taxonomies or classifications on a conceptual or theoretical foundation. For those papers that conduct reviews of existing work, most are narrative reviews, rather than the result of a systematic search (for exceptions, Hagendorff, 2024; McLean et al., 2023). In general, the state of taxonomies and classifications for AI risks are not consistent with best practice (Nickerson et al., 2013): they lack mutual exclusivity, collective exhaustivity, or parsimony; are static or based on arbitrary or ad hoc criteria; and tend to be descriptive rather than explanatory. The number of competing taxonomies inadvertently makes it very challenging to integrate relevant research into a cohesive shared understanding.

This lack of shared understanding impedes our ability to comprehensively discuss, research, and react to the risks of AI. The absence of shared understanding can cause confusion and impair research usage, cross-study comparison, and the development of cumulative knowledge (Harrison McKnight & Chervany, 2001; e.g., Marcolin et al., 2000). Shared understanding is also important in legal, political, and practical settings; reviews are frequently cited in academic and policy documents (Fang et al., 2020; Haustein et al., 2015), and shared understanding is often cited as a goal for legal and political processes (Röttinger, 2006). For example, the U.S.-EU Trade and Technology Council (TTC) stated in its joint roadmap for Trustworthy AI and Risk Management, "Shared terminologies and taxonomies are essential for operationalizing trustworthy AI and risk management in an interoperable fashion" (European Commission and the United States Trade and Technology Council, 2022).

Here, we systematically review existing AI risk classifications, frameworks, and taxonomies. We extract the categories and subcategories of risks from the included papers and reports into a "living" database that can be updated over time. We apply a "best fit" framework synthesis approach (Carroll et al., 2011, 2013) to develop two taxonomies: a high-level Causal Taxonomy of AI Risks to capture three broad causal conditions for any risk (e.g., which entities' action led to the risk, whether the risk was intentional, when it occurred), and a mid-level Domain Taxonomy which classifies the risks into seven risk domains (e.g., Discrimination and toxicity) and 23 subdomains (e.g., exposure to toxic content).

Our key contribution is the creation of a common frame of reference: an AI Risk Repository comprising a comprehensive synthesis of *existing AI risk frameworks* into a *living AI Risk Database* of 777 risks, and the Causal Taxonomy and Domain Taxonomy of AI risks. The database and taxonomies can be used individually or in combination to explore the database, as well as for research, policy, and practice to address risks from AI. All of these artifacts are available online. 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.

Methods

We used a systematic search strategy, forwards and backwards searching, and expert consultation to identify AI risk classifications, frameworks, and taxonomies. We extracted the individual risks from these documents into a living AI Risk Database. We conducted two best-fit framework syntheses to create a Causal Taxonomy of AI Risks (see Table 2) and Domain Taxonomy of AI Risks (see Table 6), by adapting existing frameworks (Weidinger et al., 2022; Yampolskiy, 2016). We did this by testing their effectiveness at coding our risk data and modifying them until we created a final version that could effectively code all relevant risks.

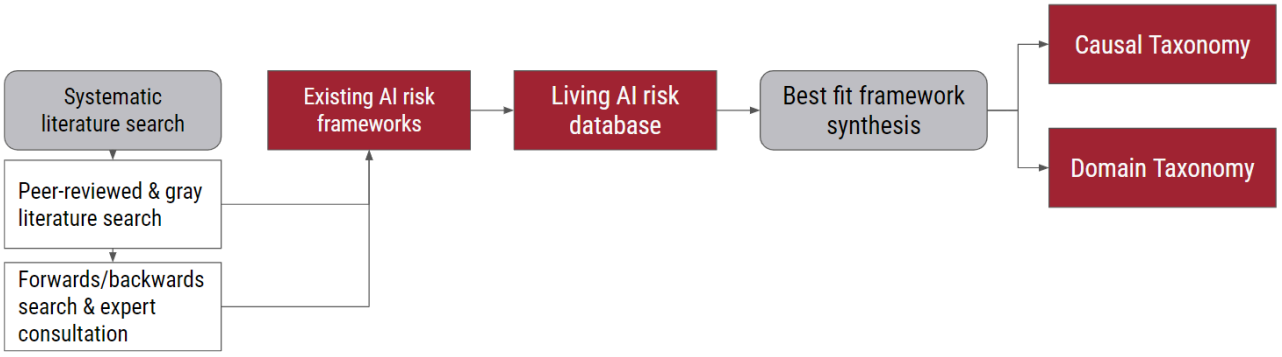


Figure 1. Overview of Study Methodology

We followed the Society for Risk Analysis (Aven et al., 2018) in defining "AI risk" as "the possibility of an unfortunate occurrence associated with the development or deployment of artificial intelligence," while recognizing that this term can be defined in many ways (Aven, 2012; Li & Li, 2023).

Systematic literature search

We conducted this study as a rapid systematic review (Khangura et al., 2012; Tricco et al., 2017). The protocol was registered in advance using the Open Science Framework in April 2024 (https://osf.io/egzc8). We included reviews, articles, reports, and documents primarily focused on proposing new frameworks, taxonomies, or other structured classifications of risks from AI present across multiple locations and industry sectors.

We excluded book chapters, theses, commentaries, editorials, and protocols. Our pilot searches suggested that including these documents would significantly increase the number of search results to screen but not the number of relevant results. We excluded documents which discussed impacts, outcomes, or other consequences of AI without specifying specific risks because we were interested in risk classification.

Due to our interest in broad, structured classifications of risks from AI, we excluded documents which focused only on risks from AI that are present in a single location or sector or that discussed risks specific to particular risk categories (e.g., content solely focused on different types of unfair decision-making) or very specific AI tools (e.g., content solely focused on risks from DALL-E). We excluded content which merely cited or discussed existing theories, frameworks, models, taxonomies, and other structured classifications rather than proposing and explaining them. This was because we wanted to understand and extract specific risks using their original source material. We excluded anything which discussed sources of risk at a high level of abstraction (e.g., the sources of sociotechnical risk in AIs) or risk-assessment processes (e.g, how organizations can assess risks from AI) rather than focusing on classifying AI risks more specifically. Non-English articles, reports, and documents were excluded due to resource constraints related to their retrieval and translation.

Two of the above exclusion criteria were added after protocol registration in order to retain only the most relevant documents: (1) Focus only on one category of risk from AI; (2) Focus on sources of risk or the risk-assessment process.

Search strategy

Our search strategy comprised two stages. In Stage 1 we conducted a systematic search of peer-reviewed and gray literature (i.e., non peer-reviewed materials) to identify relevant articles. We begin by explaining our search-term generation and strategy, followed by our database searches in Scopus and various preprint databases. We then describe our screening process, which utilizes active learning with ASReview. This process includes four phases: initial random screening for training data, application of active learning with specific stopping rules, model switching for comprehensive coverage, and quality evaluation. Finally, we outline our full-text screening and calibration procedures. In Stage 2 we conducted forwards (citation) and backwards (references) searching and expert consultation to identify additional eligible articles. The two stages are described below.

Stage 1: Searching & screening peer reviewed and gray literature

Searching

Search terms were generated through an iterative process and chosen for their empirical balance between sensitivity and specificity (Wilczynski et al., 2003). This included terms related to artificial intelligence (Artificial intelligence, AI, Artificial general intelligence, AGI), frameworks, taxonomies, and other structured classifications (Framework, Review, Overview, Taxonomy*), and risks (Risk, Harm, Hazard). This led to the following search string:

```
TITLE-ABS-KEY ( ( "artificial intelligence" OR ai OR "artificial
general intelligence" OR agi ) AND ( framework OR taxonom* OR
review ) AND ( risk OR harm OR hazard ) ) AND ( LIMIT-TO (
LANGUAGE, "English" ) ).
```

We conducted a search of Scopus to identify relevant academic research. The same search string was used on the following preprint databases to identify relevant literature: arXiv, Social Science Research Network (SSRN), Research Square, medRxiv, TechRxiv, bioRxiv, and ChemRxiv. Both searches were conducted on 4 April 2024. Relevant articles were downloaded for screening.

Title/abstract and full-text screening

Two authors formed a team to conduct title/abstract and full-text screening. Prior to screening, the team calibrated their decision-making by screening the same randomly selected articles separately (n = 23), comparing the results, and resolving disagreements. Agreement was achieved on 21 of 23 records (91%). To expedite title and abstract screening, we used a dual-screening approach with active learning in ASReview (van de Schoot et al., 2021).

Active learning is an emerging research technique which uses machine learning to reduce the total number of records that require manual screening. It is now widely used for efficiently screening large datasets in systematic reviews and meta-analyses (Campos et al., 2024; Gates et al., 2019), and has been validated in a number of diverse fields (Campos et al., 2024; van de Schoot et al., 2021) and datasets (Ferdinands et al., 2023).

Throughout the active-learning process, the four step SAFE procedure outlined by Boetje and van de Schoot (2024) was followed to ensure that screening identified relevant articles both rigorously and efficiently. The four phases are described below.

Phase 1: Screen a random set of articles to create training data for active-learning model

As per SAFE, the screening team each randomly screened and labeled 1% of the total search yield (264 records in total). Each member of the team then created separate projects in ASReview and uploaded their own files, which included all retrieved studies and the random screening data. The random screening data was automatically marked as prior knowledge, and the active-learning phase commenced.

Stage 2: Apply active learning during screening until stopping rule is reached

For the first iteration of the active-learning model, the team followed Boetje and van de Schoot's (2024) recommendation to use the Oracle model and the default model setup (TF-IDF as the feature extractor, Naive Bayes as the classifier, maximum as the query strategy and dynamic resampling (double) as the balance strategy). We aimed to follow Boetje and van de Schoot's (2024) four-fold stopping heuristics, screening until four mutually independent conditions are met:

1. All key papers are marked as relevant.
2. At least twice the estimated number of relevant records in the total dataset are screened.
3. More than 10% of the total dataset has been screened.
4. No relevant records are identified in the last 50 records screened.

These four stopping heuristics aim to achieve a sensitivity of 95% (Campos et al., 2024), ensuring comprehensive data assessment while preventing excessive time spent on unlikely candidates.

The team met three of these conditions: they i) screened more than twice the estimated number of relevant records, ii) screened more than 10% of the total dataset, and iii) had not identified any relevant records in the last 50 records. However, one condition ('All key papers are marked as relevant') could not be met due to a bug with the model. Only three out of the four key papers (Critch & Russell, 2023; McLean et al., 2023; Steimers & Schneider, 2022; Weidinger et al., 2022) had appeared in the screening process, and the final key paper was scheduled to appear several thousand papers later. Because Stage 3 of the SAFE process aims to ensure that records are not missed due to the initial model, the screening team switched models to see if a new model would locate the relevant paper.

Stage 3: Switch active-learning model and screen additional records until stopping rule is reached

Based on a review of relevant literature (e.g., Campos et al., 2024; van de Schoot et al., 2021), we use the Oracle model with the following set-up: a fully connected neural network (2 hidden layers) model as the classifier and sBert as the feature extractor, maximum as the query strategy and dynamic resampling (double) as the balance strategy. The model was trained on the data that was labeled while using the previous model. Screening stopped when no extra relevant records were identified in the last 50 records. Both authors screened in the missing key paper within the first two records found by the new model.

Stage 4: Evaluate quality

For quality checks, the screening team screened records previously labeled as irrelevant using the Oracle model and the default model set-up (i.e., the same model that was used in the initial/main model phase). This model was trained using the 10 highest- and lowest-ranked records from the model switching phase. Both team members screened records to identify any relevant records that might have been falsely excluded. This continued until the stopping rule was met (no extra relevant records identified in the last 50 records).

One member of the screening team screened the full text of all records that were included at the title/abstracts step. For calibration, 10% of the records were screened in duplicate, with 100% interrater reliability achieved. Conflicts were resolved by discussion for any remaining records.

Stage 2: Forwards and backwards searching and expert consultation

Following full-text screening, we undertook forwards and backwards searching using Scopus, Google Scholar, and various preprint servers hosting the included gray literature. Backwards searching involved identifying and reviewing all references from articles included in Stage 1, while forwards searching involved identifying and reviewing all articles that cited an included article. We also undertook an expert consultation, which involved sharing the preliminary set of included articles with their authors and other experts and requesting recommendations for relevant frameworks that had been overlooked. All records identified during forwards and backwards

searching and expert consultation were screened by one author. Those which met inclusion criteria were added to the backlog for extraction.

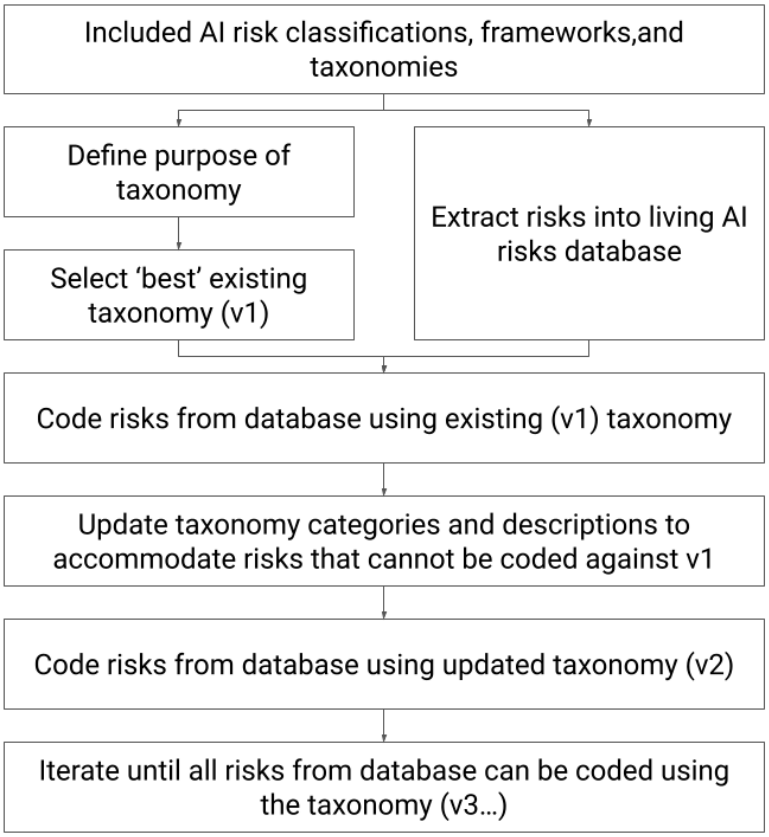
Extraction into living AI Risk Database

Five authors were involved in data extraction. A template data extraction spreadsheet was developed to capture various details from the studies, including title, abstract, author, year, source/outlet, risk category name, risk category description, risk subcategory name, risk subcategory description, and page number. This spreadsheet was refined over several rounds of pilot testing and extractor calibration on subsets of randomly selected articles. Data extraction was then conducted individually, with regular meetings for discussion and conflict resolution. Based on the recommendations of grounded theory, we aimed to capture the studied phenomena directly from the data rather than impose our interpretations (cf. Charmaz, 2006; Corbin & Strauss, 2014). Consequently, we extracted risks based on how the authors presented them, maintaining fidelity to their original categorizations and descriptions.

Best fit framework synthesis approach

Seven authors were involved in data synthesis. We used a “best fit” framework synthesis approach to develop two AI risk taxonomies. Best-fit framework synthesis is a method for rapidly, clearly, and practically understanding the relationships and structures between concepts in a topic area (Carroll et al., 2011, 2013). It combines the strengths of framework synthesis (Ritchie & Spencer, 2002), which is a “top down” positivist method where concepts are coded against a pre-existing structure, and thematic synthesis (Thomas & Harden, 2008), which is a “bottom up” interpretative method where concepts are iteratively analyzed to identify patterns and structure. We outline the process in Figure 2.

Figure 2. Methodology for Best Fit Framework Synthesis



To conduct a best-fit framework synthesis, we identified published frameworks in an area (through our systematic search and screening), selected the "best" existing framework for our purpose, then used that existing framework to code the concepts (i.e., all the risks extracted into the living AI Risk Database). Some risks could not be coded against the existing framework. We then conducted a secondary thematic analysis to identify new themes in those risks and determined what changes needed to be made to the framework to accommodate those themes. This involved updating the existing categories, creating new categories, or changing the structure of the framework. This process was repeated until achieving a final version of the framework that could most effectively code all relevant risks.

By starting with an existing framework, the synthesis can achieve a coherent framework more quickly than inductively or thematically analyzing all the individual concepts (risks) across all included papers. The trade-off is that the existing framework creates a particular "lens" for understanding and categorizing the individual concepts, which may lead to a disconnect between the synthesized findings and the theoretical or epistemological perspectives in the original and highly varied papers. In order to mitigate this, we attempted to code the risks based on the exact wording the authors had presented rather than our interpretation of what they may have intended to communicate (cf. Charmaz, 2006; Corbin & Strauss, 2014).

Why we developed two taxonomies of AI risk

The goal of our best-fit framework synthesis was to create a common frame of reference for understanding and addressing the risks from artificial intelligence. We found that authors implicitly

or explicitly used different lenses (e.g., Head, 2008; Nilsen, 2015; Sovacool & Hess, 2017) to create their frameworks. These lenses reveal and obscure different aspects of the AI risk landscape.

Through our systematic search, we identified two types of frameworks, which we will refer to here as high- and mid-level frameworks. “High-level frameworks” focused on capturing broad factors that specify how, when, or why an AI risk might emerge (e.g., Critch & Russell, 2023; Kilian et al., 2023) rather than discuss categories of specific hazards and harms. In contrast, “Mid-level frameworks” focused on specific hazards and harms (e.g., Solaiman et al., 2023; Weidinger et al., 2021) but didn’t explore their causes.

The fact that categorizations were at such different levels of specificity made it challenging to create a single framework. Often, specific mid-level risks did not fit into the categories within a high-level framework, and the broad categories in those frameworks were insufficiently specified to be useful, in isolation, for creating shared understanding. Similarly, broad high-level categorizations of how, when, or why an AI risk emerges did not fit with mid-level frameworks outlining narrow and specific sets of risk.

We therefore resolved that the ideal common frame of reference required two intersecting taxonomies: one to precisely decompose or define an AI risk based on the high-level conditions under which it occurred (a “causal taxonomy”), and one that classified commonly discussed hazards and harms associated with AI into understandable and distinct domains (a “domain taxonomy”).

In the following sections, we describe the process of developing these two taxonomies using a best-fit framework synthesis approach.

Development of high-level Causal Taxonomy of AI Risks

Best-fit taxonomy: Yampolskiy (2016)

We chose Yampolskiy’s (2016) *Taxonomy of pathways to dangerous AI* as our initial best-fit framework for developing a causal taxonomy for AI risk. We selected Yampolskiy’s taxonomy as it was highly cited (116 citations, fifth most highly cited from the set of identified papers), simple, comprehensive, and provided sufficient definitions for each category.

Yampolskiy’s taxonomy systematically classifies the ways in which an AI system might become dangerous based on two main factors: Timing - whether the AI became dangerous at the pre-deployment or post-deployment stage, and Cause - whether the danger arose from External Causes (On Purpose, By Mistake, Environment) or Internal Causes originating from the AI system itself (Independently).

Yampolskiy proposes that this taxonomy covers scenarios ranging from AI being purposely designed to be dangerous, to becoming dangerous by accident during development or after deployment, to turning dangerous due to environmental factors outside its control, or evolving to become dangerous through recursive self-improvement. Each “pathway” represents a set of causal conditions that lead to AI causing harm, e.g., a person using a large language model (LLM) to generate fake news for political gain would be classified under Path B (“Timing: post-deployment; External cause: on purpose”).

Coding and iteration process

We started by using Yampolskiy's taxonomy to categorize a sample of risks from our database. We then identified themes in the AI Risk Database that didn't fit into Yampolskiy's taxonomy. We updated the taxonomy categories, criteria, and descriptions, then coded a further sample of risks. This process was repeated over three iterations until the taxonomy categories, criteria, and descriptions were stable. We describe this iteration process in detail in Appendix A.

Final taxonomy

The final version of the taxonomy, which we named the ***Causal Taxonomy of AI Risks***, included three categories of causal factors that specify how, why, or when an AI risk might emerge. The first category, ***Entity***, classified the entity (e.g., AI system, Human) that was presented as causing the risk to occur due to a decision or action taken by that entity. The second category, ***Intent***, classified whether the risk was presented as an expected outcome or unexpected outcome of an entity pursuing a goal. The third category, ***Timing***, classified the stage in the AI lifecycle that the risk is presented as occurring (e.g., pre-deployment, post-deployment). Each of these categories includes a third option, "Other," which captures risks that are not clearly categorizable within the primary options. Each of the categories is therefore mutually exclusive; each risk is classified under only one option within each category. The Causal Taxonomy is presented and described in more detail in the Results section.

Development of mid-level Domain Taxonomy of AI Risks

Best-fit taxonomy: Weidinger (2022)

We chose Weidinger et al (2022) “Taxonomy of Risks posed by Language Models” as our initial mid-level best-fit framework because it and its related papers (Weidinger et al., 2021, 2023) are among the highest cited in our review. Although this taxonomy was focused on language models, its set of categories was one of the most comprehensive, and it has been iterated upon over several publications. It included six areas of risks from language models: (1) Discrimination, Hate speech and Exclusion; (2) Information Hazards; (3) Misinformation Harms; (4) Malicious Uses; (5) Human-computer interaction Harms; and (6) Environmental and Socioeconomic Harms. Each area of risk described several subcategories of risk.

Coding and iteration process

We applied this taxonomy by coding as many of the included risks as possible using the Weidinger (2022) taxonomy. We operationalised the taxonomy by using the definitions or descriptions for each category from Weidinger (2022). Because several similar taxonomies were included in the set identified by the systematic literature review (e.g., Weidinger et al., 2021, 2022, 2023), we considered descriptions and definitions from any of these taxonomies in our initial coding.

We iterated on the taxonomy to accommodate risks that could not be coded against the existing Weidinger (2022) taxonomy. The most common risks that could not be accommodated were those related to AI system safety, failures and limitations; AI system security vulnerabilities and attacks;

and competitive dynamics or other failures of governance to manage the development and deployment of AI systems. We describe this iteration process in detail in Appendix A.

Final taxonomy: Domain Taxonomy of AI Risks

The final version of the taxonomy, which we named the Domain Taxonomy of AI Risks, included seven domains of AI risk, and 23 subdomains of hazards and harms associated with AI. The domains were (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***. As with Weidinger's (2022) taxonomy, these risk domains are not mutually exclusive; many risks span multiple domains or subdomains due to their interconnected nature. For example, a risk related to AI-generated disinformation could be relevant to both the *Misinformation* domain and the *Malicious actors & misuse* domain. The Domain Taxonomy is presented and described in more detail in the Results section.

Coding

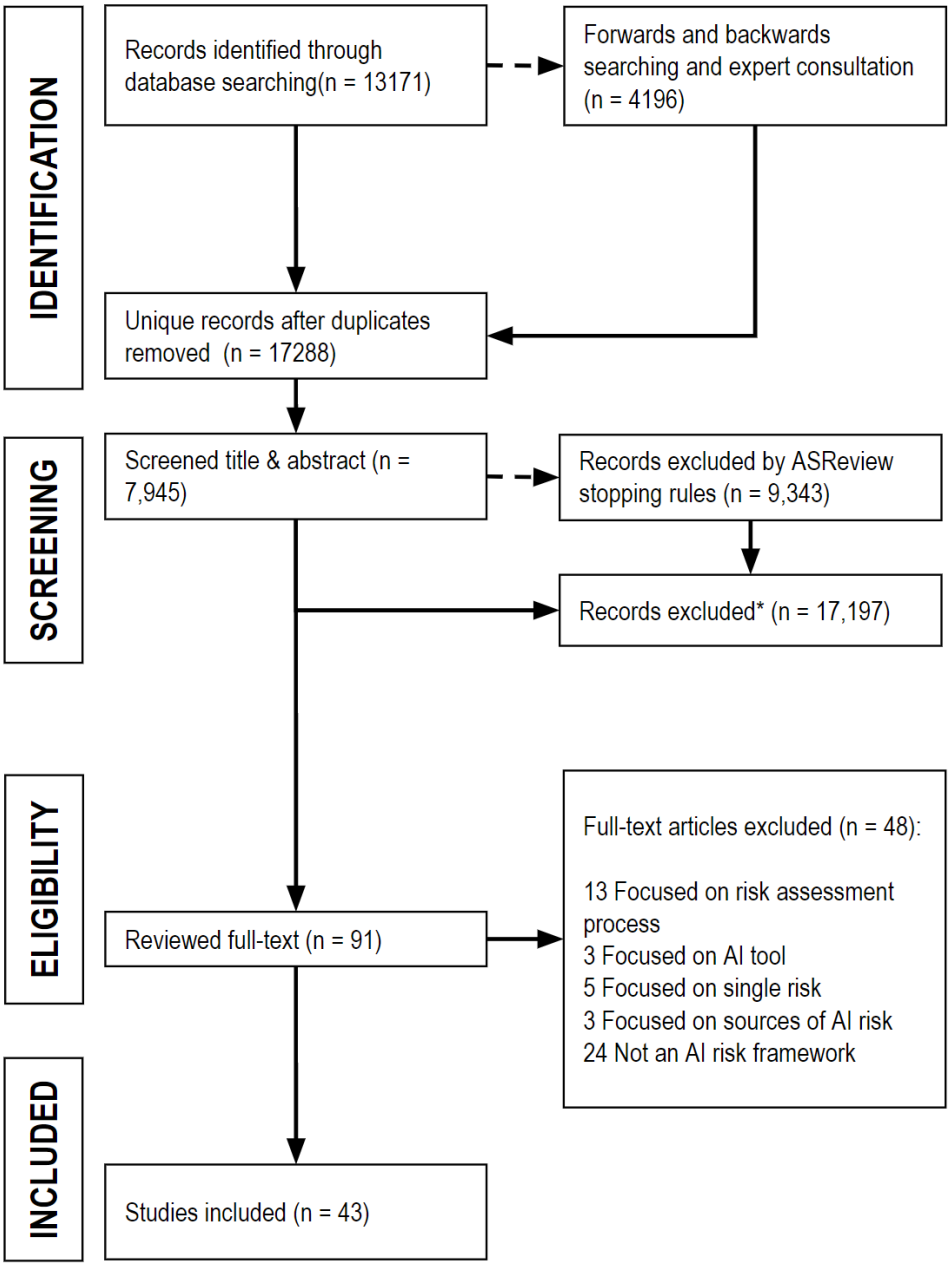
Three authors were involved in coding risks against our taxonomies. Risks were coded by a single reviewer and discussed with the team where relevant. The coding process involved systematically categorizing each extracted risk according to the definitions within the relevant taxonomy. Based on grounded theory recommendations (cf. Charmaz, 2006; Corbin & Strauss, 2014), we coded risks as they were presented by the authors, aiming to capture the studied phenomena directly rather than impose our own interpretations or infer intent. When coding risks for our high-level Causal Taxonomy, we categorized risks relevant to multiple levels of each causal factor (e.g., both pre-deployment and post-deployment) as "Other." In our mid-level Domain Taxonomy, we categorized risks relevant to multiple domains and subdomains (e.g., AI-generated disinformation) in the single most relevant category.

Results

Systematic literature search

We retrieved 17,288 unique articles from our searches and expert consultations. Of these records we screened 7,945. We excluded 9,343 records were not screened by the authors; they were excluded by our stopping criteria while using ASReview which used machine learning to recommend when screening was unlikely to yield further relevant content. We assessed the full text of 91 articles. A total of 43 articles and reports met the eligibility criteria: 21 from our search, 13 from forwards and backwards searching, and 9 from expert suggestions. We present a PRISMA diagram illustrating the search results.

Figure 3. PRISMA Flow Diagram for Systematic Literature Search and Screening



*Note. *The ASReview software was used to assist in screening titles and abstracts.*

Characteristics of included documents

We included 43 documents: 17 peer-reviewed articles, 16 preprints, 6 conference papers, and 4 other reports. We mainly identified recent literature, with all but four of the included documents published later than 2020.

We coded the corresponding author's country and found that most of the included documents were from the USA (n = 12), China (n = 8), United Kingdom (n = 6), and Germany (n = 5). Other countries included Singapore, Portugal, Australia, Brazil, India, Iran, Netherlands, Spain, and Türkiye. Most of the included documents had a corresponding author affiliated with a university (n = 26), followed by industry organization (n = 11), with the remainder from government, international organizations (i.e., United Nations), or non-government organizations (n = 6). The most common

affiliation was DeepMind / Google DeepMind (n = 5). The most common described methodology was a narrative review or "survey" of existing literature (n = 16), followed by systematic review (n = 6), and scoping review (n = 2). Most of the papers did not explicitly describe their methodology in any detail.

The included documents varied in the type or scope of Artificial Intelligence they focused on. In most cases, the type of AI was not explicitly defined (n = 19). Large language model was the next most common (n = 10), followed by Generative AI (n = 6), Machine Learning (n = 2), and Artificial General Intelligence (n = 2). Other terms included "AI and Machine Learning" (always described in the document as "AI/ML"), AI assistant, algorithmic systems, and general-purpose AI.

The framing of risk and AI risk differed significantly across documents. Only three documents explicitly defined risk, describing it as "the impact of uncertainty on objectives" (Steimers & Schneider, 2022), "the consequence of an event combined with its likelihood of occurrence" (Tan et al., 2022), or as a notional equation in conjunction with a definition of a hazard: "A Hazard is a source of danger with the potential to harm… Risk = Hazard × Exposure × Vulnerability" (Hendrycks and Mazeika 2022). The classifications, frameworks, and taxonomies used varied terms to describe risks, including: "risks of/from AI," "harms of AI," "AI ethics," "ethical issues/concerns/challenges," "social impacts/harms," and others.

Together, the included documents presented a total of 777 risk categories (e.g., "Privacy risks") or sub-categories (e.g., "Compromising privacy by leaking sensitive information", Weidinger et al., 2022) of risk. Not all documents presented a framework with both categories and sub-categories. Not all documents presented eligible risk categories in sufficient detail to allow us to code them with our taxonomies; two included documents were not coded as having any distinct risk categories or framework (AI Verify Foundation, 2023; e.g., Sharma, 2024). These are therefore unrepresented in later outputs. Our supplementary online resources include a database of all risks and included documents & frameworks. The full set of included documents are also presented in Appendix B.

The included documents contained a range of highly cited taxonomies, as shown in Table 1.

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.

Causal Taxonomy of AI Risks

As a result of a best-fit framework synthesis, which selected, adapted, and iterated on a high-level taxonomy of causal factors (Yampolskiy, 2016) identified in our systematic search, we developed the Causal Taxonomy of AI Risks (Table 2). This taxonomy uses Entity, Intent, and Timing to classify risks in the AI Risk Database. We coded 668 of the 777 (86%) risks extracted from our documents against the Causal Taxonomy. 65 did not present sufficient information to assess the Entity, Intent, or Timing, and 43 were discarded as they did not fit our definition of risk.

Table 2. 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

The ***Entity*** variable captures which, if any, entity is *presented* as the main cause of the risk. It includes three levels: AI, Human, and Other. When the risk is attributed to AI, it means that the risk arises from decisions or actions made by the AI system itself, such as generating harmful content or disempowering humans. Conversely, when humans are seen as the source, the risks are implied to be due to human actions like choosing poor training data, intentional malicious design, or improper use of AI systems. The "Other" category captures cases where the focal entity is not a human or AI or is ambiguous. For example, “The software development toolchain of LLMs is complex and could bring threats to the developed LLM,” implies that the toolchain could be exploited by humans or AI.

The ***Intent*** variable captures whether the risk is *presented* as occurring as an expected or unexpected outcome from pursuing a goal. This variable has three levels: Intentional, Unintentional, and Other. Intentional risks are those that occur as expected outcomes from pursuing a specific goal, such as a case where AI is intentionally programmed to act deceptively or to exhibit bias. Unintentional risks reflect unintended consequences, such as an AI system inadvertently developing biases due to incomplete training data. The "Other" category captures risks where the intent is not clearly specified; for example, “The external tools (e.g., web APIs) present trustworthiness and privacy issues to LLM-based applications.” This includes cases where the risk may occur intentionally and unintentionally, such as "The potential for the AI system to infringe upon individuals' rights to privacy, through the data it collects, how it processes that data, or the conclusions it draws."

The ***Timing*** variable captures the stage in the AI lifecycle at which the risk is presented as occurring. The levels within this variable include Pre-deployment, Post-deployment, and Other. Pre-deployment risks are those that arise before the AI system is fully developed and put into use,

such as vulnerabilities in the model due to coding errors. Post-deployment risks arise after the AI has been deployed, including issues like the misuse of AI for harmful purposes. Deployment is not defined in Yampolskiy (2016); we therefore interpreted it to mean when a product is being used by end users rather than just by developers. The "Other" category is used for risks that do not have a clearly defined time of occurrence (e.g., ”Resilience against adversarial attacks and distribution shift”). This includes cases where the presented risk may occur both before and after deployment; for example, "Generative models are known for their substantial energy requirements, necessitating significant amounts of electricity, cooling water, and hardware containing rare metals.”

Most common causal factors for AI risk

Table 3 shows how the risks were coded against each category of causal factors. A majority of the risks were presented by authors of the documents as due to a decision or action by an artificial intelligence system (51%). Slightly more risks were presented as Unintentional (37%) compared to intentional (35%). Most of the risks were presented as occurring post-deployment (65%).

Table 3. 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.

Table 4 shows how the risks intersect across our three causal factors. The most common triads of causal conditions under which an AI risk was presented as occurring were Entity = AI, Intention = Unintentional, Timing = Post-deployment (18% of all risks). This was followed by Entity = Human, Intention = Intentional, Timing = Post-deployment (17% of all risks).

Table 4. AI Risk Database coded with Causal Taxonomy: entity x intent x timing

		Intent		
Timing	**Entity**	*Intentional*	*Unintentional*	*Other*
Pre-deployment	*Human*	2%	•	•
	AI	•	3%	•
	Other	•	•	•
Post-deployment	*Human*	17%	4%	2%
	AI	4%	18%	11%
	Other	2%	2%	3%
Other	*Human*	4%	•	•
	AI	4%	6%	2%
	Other	•	•	4%

Note. Taxonomy categories with a prevalence ≤ 10% are highlighted. Categories with a prevalence less than 2% in the AI Risk Database are shown as • for ease of interpretation.

Causal factors of AI risk examined by included documents

Papers varied significantly in terms of which causal factors they examined (see Table 5). Two documents have blank rows in the table because they did not present risks that could be coded against the Causal Taxonomy (#26, AI Verify Foundation, 2023; #36, Sharma, 2024). Human-related risks were identified in 34 out of 41 documents (83%), AI-related risks in 40 out of 41 documents (98%), and other risks in 28 out of 41 documents (68%). Regarding intent, intentional risks were noted in 35 out of 41 documents (85%), unintentional risks in 35 out of 41 documents (85%), and other intent-related risks in 36 out of 41 documents (88%). In terms of timing, pre-deployment risks were identified in 22 out of 41 documents (54%), post-deployment risks in 37 out of 41 documents (90%), and other timing-related risks in 33 out of 41 documents (80%). The majority of documents recognized risks arising from both human and AI actions, with equal acknowledgment of intentional and unintentional risks. Post-deployment risks were more frequently discussed than pre-deployment risks, indicating the documents included focused more on the consequences of deployed AI systems.

Table 5. Included documents coded with causal taxonomy

		Entity			Intent			Timing		
ID	First Author (Year)	*Human*	*AI*	*Other*	*Intentional*	*Unintentional*	*Other*	*Pre-deployment*	*Post-deployment*	*Other*
1	Critch (2023)	X	X		X	X			X	X
2	Cui (2024)	X	X	X	X	X	X	X	X	X
3	Cunha (2023)	X	X	X	X	X	X		X	X
4	Deng (2023)		X	X	X	X	X	X	X	
5	Hagendorff (2024)	X	X	X	X	X	X	X	X	X
6	Hogenhout (2021)	X	X		X	X	X	X	X	X
7	Kilian (2023)	X	X	X	X	X	X		X	X
8	McLean (2023)	X	X	X			X	X	X	X
9	Meek (2016)	X	X	X	X	X	X		X	X
10	Paes (2023)	X	X	X	X	X			X	
11	Shelby (2023)	X	X	X	X	X	X		X	X
12	Sherman (2023)	X	X	X	X	X	X		X	X
13	Solaiman (2023)	X	X	X	X	X	X	X	X	X
14	Steimers (2022)		X	X		X	X		X	X
15	Tan (2022)	X	X	X	X	X	X	X	X	X
16	Weidinger (2022)	X	X	X	X	X	X		X	X
17	Weidinger (2021)	X	X	X	X	X	X		X	X
18	Weidinger (2023)	X	X	X	X	X	X	X	X	X
19	Wirtz (2022)	X	X	X	X	X	X	X	X	X
20	Wirtz (2020)	X	X	X	X	X	X		X	X
21	Zhang (2022)	X	X	X	X	X		X		X
22	Hendrycks (2023)	X	X		X	X	X	X	X	X
23	Vidgen (2024)		X				X		X	
24	Gabriel (2024)	X	X	X	X	X	X	X	X	X
25	Shevlane (2023)		X		X			X	X	X
26	AIVerify (2023)									
27	Sun (2023)	X	X		X		X		X	
28	Zhang (2023)		X			X	X		X	
29	Habbal (2024)	X	X		X	X			X	
30	Liu (2024)	X	X		X	X	X	X	X	X
31	EPIC (2023)	X	X	X	X	X	X	X	X	X

		Entity			Intent			Timing		
ID	First Author (Year)	*Human*	*AI*	*Other*	*Intentional*	*Unintentional*	*Other*	*Pre-deployment*	*Post-deployment*	*Other*
32	Stahl (2024)		X				X			X
33	Nah (2023)	X	X	X	X	X	X	X	X	X
34	Ji (2023)	X	X	X	X	X	X	X	X	X
35	Hendrycks (2022)	X	X	X	X	X	X	X	X	X
36	Sharma (2024)									
37	Giarmoleo (2024)	X		X	X		X	X		X
38	Kumar (2023)		X			X	X		X	X
39	Saghiri (2022)	X	X		X	X	X	X	X	X
40	Yampolskiy (2016)	X	X	X	X	X	X	X	X	
41	Allianz (2018)	X	X	X	X	X	X		X	
42	Teixeira (2022)	X	X	X	X	X	X	X	X	X
43	InfoComm (2023)	X	X		X	X	X			X

Domain Taxonomy of AI Risks

As a result of a best-fit framework synthesis, which selected, adapted, and iterated on a mid-level taxonomy of categories of AI risk (Weidinger et al., 2022), we developed the Domain Taxonomy of AI Risks. This taxonomy catalogs hazards and harms associated with AI and was used to classify risks in the AI Risk Database. We coded 706 (91%) of the 777 risks extracted from our documents against the Domain Taxonomy. 71 risks did not present sufficient information for coding.

In Table 6, we present the Domain Taxonomy including subdomains and short descriptions of each subdomain. In the following sections we present a detailed description of each subdomain using information from included documents in the AI Risk Database, as well as analysis of the distribution of risks and domains across the database.

Table 6. 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 unfair 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 is 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, resulting 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 such as 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 (e.g., medical emergency). Over reliance 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, potentially leading 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, including 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, such as 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.

Detailed descriptions of domains of AI risks

Domain 1: Discrimination and toxicity

1.1 Unfair discrimination and misrepresentation. Humans hold inaccurate and overgeneralized beliefs about the characteristics, behaviors, and attributes of members of certain social groups. These stereotypical beliefs and the behavior that follows from them can misrepresent, exclude, demean, and disadvantage the individuals to whom they apply, reinforcing existing inequality. Human belief and behaviors shape every part of the design, development, and deployment of AI. Humans program AI systems, provide training data, and decide how data is processed and stored (Wirtz et al., 2022). As a result, AI models can encode associations that promote and amplify biased or discriminatory beliefs and behaviors. In decision systems, erroneous associations can systematically disadvantage certain groups. This may result in harmful decisions such as wrongful rejection of loan or mortgage applications (Shelby et al., 2023; Wirtz et al., 2020), discriminatory hiring practices that exclude qualified candidates (Kumar & Singh, 2023; Shelby et al., 2023; Wirtz et al., 2020), or the misidentification and unjust arrest of individuals in law enforcement contexts (Kumar & Singh, 2023; Paes et al., 2023). In text and image models, biased inputs can manifest in outputs that reinforce harmful stereotypes and prejudices that paint certain groups and individuals "... as lower status and less deserving of respect" (Shelby et al., 2023).

1.2 Exposure to toxic content. Certain types of content have the potential to cause harm to the people who are exposed to them. These harms can vary in impact from minor (e.g., a transient experience of discomfort) to more severe (e.g., psychological, social, or physical consequences that are significant and/or enduring). Harmful speech is prevalent on the internet, particularly on social media platforms (Castaño-Pulgarín et al., 2021). Because AI models are commonly trained on vast amounts of internet data, they can internalize and regenerate these speech patterns in their output. In the context of LLMs, this output is known as "toxic content," an umbrella term that includes harmful, abusive, unsafe, and offensive material that violates community standards (Infocomm Media Development Authority, 2023; Shelby et al., 2023). Frequently observed categories include content that promotes or encourages unlawful activities, hate, extremism, and violence (Cui et al., 2024; Hagendorff, 2024; Vidgen et al., 2024; Weidinger et al., 2023); provides hazardous or misleading high-risk advice (Hagendorff, 2024; Vidgen et al., 2024); or contains unwelcome or profoundly offensive, explicit material such as profanity, pornography, or child sexual abuse imagery (Liu et al., 2023; Weidinger et al., 2023).

1.3 Unequal performance across groups. Decisions made during the development of an algorithmic system and the content, quality, and diversity of the training data can significantly impact which people and experiences the system can effectively understand, represent, and accommodate (Shelby et al., 2023; Solaiman et al., 2023). Biases and limitations introduced through these factors can lead to models that perform significantly worse for certain subpopulations compared with others, especially those defined by disability, gender identity, race, social status, and ethnicity (Liu et al., 2023; Shelby et al., 2023). For example, when LLMs are trained on a small number of languages, they can underperform for others (Weidinger et al., 2021, 2022). The underperformance of algorithmic systems for certain groups may lead to a range of negative consequences such as the reduced ability or complete inability to use and benefit from the system (Shelby et al., 2023); increased effort or challenges in using it effectively (Shelby et al., 2023); feelings of alienation,

frustration, and exclusion due to the lack of inclusive design (Shelby et al., 2023); and ultimately, unequal outcomes across various domains (Shelby et al., 2023; Solaiman et al., 2023).

Domain 2: Privacy & security

2.1 Compromise of privacy by obtaining, leaking, or correctly inferring sensitive information. In the context of generative AI, privacy violations arise when systems collect and divulge sensitive information that individuals or corporations do not consent to sharing with others (Cui et al., 2024; Hagendorff, 2024; Sherman & Eisenberg, 2024; Steimers & Schneider, 2022; Vidgen et al., 2024; Weidinger et al., 2023). Privacy violations can occur both accidentally and intentionally.

Examples of accidental causes include AI models that memorize and inadvertently reproduce or leak sensitive personal information present in their training data, such as names, addresses, and medical records (Cui et al., 2024; Deng et al., 2023; Hagendorff, 2024; Shelby et al., 2023; Weidinger et al., 2021). Even when personal data is not included in the training dataset or directly offered by the user, models can make inferences about sensitive or protected traits of individuals based on predictive correlations within their history of interactions (Cunha & Estima, 2023; Weidinger et al., 2021), build profiles of users (Hogenhout, 2021; Shelby et al., 2023), or train AI systems (Habbal et al., 2024). As a result, models may save and reproduce sensitive information derived from prior interactions, such as classified intellectual property (Cunha & Estima, 2023; Weidinger et al., 2021). A notable example is the case where Samsung employees accidentally leaked confidential intellectual property to OpenAI after using ChatGPT to help with coding tasks (Cunha & Estima, 2023; Weidinger et al., 2021).

Intentional causes include the malicious design and use of AI to exploit users' trust by influencing them to share personal or private information about themselves or others (Gabriel et al., 2024). Privacy attacks, such as membership inference, could allow adversaries to gain knowledge of the private records used to train an AI model (Gabriel et al., 2024). Malicious actors could also deliberately extract private information from a model by crafting prompts designed to exploit the model's knowledge of sensitive data (Cui et al., 2024).

2.2 AI system security vulnerabilities and attacks. AI systems, like other software systems, face a range of security threats. These issues may arise from inherent weaknesses in the design of AI algorithms, the data used to train the models, or the operational context. Specific examples include:

- *Toolchain and dependency vulnerabilities* that arise unintentionally through the use of automated code-generation tools (e.g., Github Copilot, Python language, OpenCV), deep learning frameworks (e.g., Tensorflow, PyTorch), or as a result of complex interdependencies in the development environment (Cui et al., 2024).
- *External tool and API integration into AI system applications* can compromise the trustworthiness and privacy of systems due to their potential unreliability or susceptibility to adversarial control (Cui et al., 2024).
- *Security vulnerabilities in physical and network infrastructure,* such as vulnerabilities in graphics processing units, or GPUs, or to sophisticated attacks like side-channel and rowhammer attacks, can lead to unauthorized access or manipulation of model parameters when used during training of AI systems (Cui et al., 2024). The use of distributed network systems for training AI systems such as LLMs exposes them to network-specific threats like pulsating attacks or congestion.

- *Direct manipulation of AI systems* such as adversarial attacks and instruction-based attacks. Adversarial attacks focus on altering the model's learning process or extracting its data. They include perturbations designed to deceive models into incorrect outputs, extraction attacks to steal model insights, and poisoning attacks to alter model behavior (Cui et al., 2024; Gabriel et al., 2024; Liu et al., 2023). Instruction-based attacks manipulate the way the model handles and responds to inputs (Hagendorff, 2024; Liu et al., 2023; Sun et al., 2023). Attackers deliberately craft prompts to induce models to produce biased or unsafe outputs (a.k.a. 'jailbreaking'). This manipulation directly targets the operational aspects of AI systems with the intent to cause harm.

Domain 3: Misinformation

3.1 False or misleading information. LLMs can sometimes generate content that is factually incorrect, misleading, poorly researched, or unintelligible (Cunha & Estima, 2023; Deng et al., 2023; Electronic Privacy Information Centre, 2023; Hagendorff, 2024; Nah et al., 2023; Shelby et al., 2023). Risks in this category occur accidentally and not as a result of humans intentionally trying to cause harm, as is the case with *disinformation (Liu et al., 2023; Weidinger et al., 2022)*. Common sources of AI misinformation include noisy training data (Cui et al., 2024; Liu et al., 2023), sampling strategies that introduce randomness (Cui et al., 2024), outdated knowledge bases (Liu et al., 2023), and fine-tuning processes that encourage sycophantic behavior (Cui et al., 2024). Incorrect and misleading information generated by LLMs can result in a range of actual and anticipated negative outcomes. Individuals exposed to false information may form inaccurate beliefs and perceptions. This undermines their autonomy and ability to make free and informed choices (Weidinger et al., 2021, 2022). Where inaccuracies in LLM predictions influence an individual's decisions and actions, the individual may experience indirect physical, emotional, or material harms (Gabriel et al., 2024; Weidinger et al., 2022) especially but not exclusively in high-stakes domains such as mental health (Sun et al., 2023; Z. Zhang et al., 2023), physical health (Nah et al., 2023; Sun et al., 2023; Z. Zhang et al., 2023), law (Weidinger et al., 2022), and finance (Vidgen et al., 2024). For example, an LLM that offers misleading information about medical drug use may cause a consumer to harm themselves or others (Sun et al., 2023).

3.2 Pollution of information ecosystems and loss of consensus reality. This subcategory covers the diverse effects of AI-driven personalisation and content-generation technologies on the information landscape. As AI systems become more adept at tailoring content to individual preferences, they risk creating "filter bubbles" (Hogenhout, 2021). These are informational cocoons where individuals are predominantly exposed to news and opinions that align with their pre-existing beliefs. AI-driven filter bubbles are likely to be more pervasive and intense than those driven by traditional internet browsing and recommendation algorithms: They adapt to individual preferences in a more sophisticated manner (e.g., through reinforcement learning and analysis of user behavioral data) (Gabriel et al., 2024), integrate seamlessly into daily life, and are more opaque. An overreliance on hyper-personalized AI information sources could lead to a "splintering" of shared reality, where different groups of people have vastly different understandings of what is true or important (Gabriel et al., 2024; Hogenhout, 2021). This is likely to be exacerbated by the proliferation of AI-enabled content generation technologies that spread misinformation at higher rates (e.g., clickbait), potentially making consumers generally distrustful of information and important institutions (Electronic Privacy Information Centre, 2023; Gabriel et al., 2024; Hendrycks & Mazeika, 2022). A shared sense of reality is fundamental to social solidarity. Where societal

bonds are weakened, individuals may become more hostile towards opposing views. This can hinder constructive dialogue on critical collective issues like climate change and public health (Gabriel et al., 2024).

Domain 4: Malicious actors and misuse

4.1 Disinformation, surveillance, and influence at scale. Advances in AI have made powerful dual-use technologies like voice cloning, deep fakes, content generation, and data-gathering tools cheaper, more efficient, and easier to use (Cunha & Estima, 2023). With modest hardware requirements, these technologies are now within the reach of a broader group of users, including those with malicious intent. Disinformation is already a serious issue (Hendrycks et al., 2023) and involves the deliberate propagation of false or misleading information, usually with the intent to cause harm, influence behavior, or achieve a financial or political advantage (Electronic Privacy Information Centre, 2023; Infocomm Media Development Authority, 2023).

AI tools could be used to amplify the impact and scope of disinformation through more personalized, convincing, and far-reaching messaging (Gabriel et al., 2024; Hendrycks et al., 2023; Weidinger et al., 2021, 2022; Wirtz et al., 2020). For example, the use of advanced AI in phishing schemes enables cybercriminals to automate the creation of highly sophisticated image, video, and audio communications (Cunha & Estima, 2023; Gabriel et al., 2024; Habbal et al., 2024). These communications can be tailored to individual recipients (sometimes including the cloned voice of a loved one), making them more likely to be successful and harder for both users and anti-phishing tools to detect (Gabriel et al., 2024). In the realm of surveillance, AI could support and enhance the mass gathering of personal data (Gabriel et al., 2024; Weidinger et al., 2021, 2022). Historically, mass surveillance required extensive manual effort. Machine learning tools can now link and process large datasets much more efficiently and cheaply than human analysts and can make predictions and decisions without human intervention (Weidinger et al., 2022). Through microtargeting, actors could manipulate individual behavior more subtly and effectively using AI-derived insights from their personal data and online behavior.

In the hands of nefarious state actors, such capabilities could be used to enhance the effectiveness of illegitimate domestic surveillance campaigns and to facilitate oppression and control (Gabriel et al., 2024). All of the capabilities mentioned above could converge to facilitate the large-scale manipulation and control of what people see, hear, and believe. A form of this is automated censorship in which AI systems are used to selectively suppress or block specific types of information, content, or voices deemed undesirable to those controlling the AI (Wirtz et al., 2022). AI can not only be used to silence voices but also to entrench specific agendas: Actors (be they political figures, organizations, or state actors) could use AI to distribute incorrect information about electoral systems and processes (Vidgen et al., 2024), produce persuasive propaganda (Allianz Global Corporate & Security, 2018; Liu et al., 2023), and systematically exert control over public opinion and political debates on a large scale (Wirtz et al., 2022). During the 2016 Brexit referendum, a network of over 10,000 AI-powered political bots were employed to distribute fake and hyperpartisan news (Allianz Global Corporate & Security, 2018; Bastos & Mercea, 2019). The selective visibility of information can lead to the formation of incorrect or incomplete beliefs about what is happening in the world. This ability to shape public discourse can maintain or increase the power of those in control while keeping the public in the dark about critical issues that may affect their lives and their society.

4.2 Cyberattacks, weapon development or use, and mass harm. AI may be used to gain a political or strategic advantage or to cause harm at scale through cyber operations or the development and use of weapons. Advancements in AI have provided malicious actors with powerful tools that can lead to more frequent, more severe, and more precise cyber attacks (Gabriel et al., 2024; Hendrycks & Mazeika, 2022; Hogenhout, 2021; Kilian et al., 2023; Liu et al., 2023; Weidinger et al., 2022; Wirtz et al., 2022). Hackers could use the coding abilities of AI assistants to develop malicious malware more effectively and at lower cost (Allianz Global Corporate & Security, 2018; Hagendorff, 2024; Hogenhout, 2021; Sherman & Eisenberg, 2024; Weidinger et al., 2022). With AI, even those with limited coding and technical experience (Electronic Privacy Information Centre, 2023; Gabriel et al., 2024) could teach a model to produce and optimize malware code that discovers and exploits system vulnerabilities, including both self-replicating (Allianz Global Corporate & Security, 2018; Infocomm Media Development Authority, 2023) and automated software (Cui et al., 2024; Shevlane et al., 2023). The development and application of weapons could also be sped up and intensified through AI. For example, AIs with specialized knowledge of bioengineering could make it easier for more actors to design new bioweapons (Hendrycks et al., 2023; Infocomm Media Development Authority, 2023). For example, in 2022 a small pharmaceutical company used generative AI to develop 40,000 chemical nerve agents in less than six hours (Sohn, 2022). AI could also enable autonomous devices, such as drones, to be used as weapons (Allianz Global Corporate & Security, 2018). In fact, AI has already assisted in the development and application of Lethal Autonomous Weapons Systems (LAWS) – weapons that can operate without human oversight and use computer algorithms to identify and attack targets (Habbal et al., 2024; Hogenhout, 2021). Autonomous weapons may fail in ways that other AI systems do, such as through a lack of capability, robustness, or loss of control, meaning that they would cause harm that was not intended by their developers or operators. In these circumstances, the risks from LAWS would not be limited to malicious actors or misuse. However, in most cases, we conceptualize risks from LAWS as relating to purposeful decisions made by humans controlling the weapons. AIs deployed by states in conflict could also be integrated in conventional defense or mass-casualty weapons (Teixeira et al., 2022; Vidgen et al., 2024). These integrations could range from AI-controlled aerial combat to the operation of AI as part of a country's nuclear arsenal as a “fail-safe” mechanism (Hendrycks & Mazeika, 2022).

Overall, AI’s ability to process vast amounts of data quickly may empower actors to act on a much larger scale than would otherwise be possible. AI can manage multiple attack vectors simultaneously, coordinating them to maximize disruption and harm. Malicious actors may intentionally cause mass harm through terrorism or the disruption of law enforcement (Critch & Russell, 2023). For example, AI could automate the process of finding and exploiting vulnerabilities in software used by millions of people (Gabriel et al., 2024). AI could also be used to identify vulnerabilities in national power grids and strategically target key components to cause outages or to determine optimal release points for biological agents to maximize impact and spread.

4.3 Fraud, scams, and targeted manipulation. AI capabilities have the potential to be exploited for personal gain at the expense of others via deception and manipulation. This can take various forms including cheating, fraud, scams, and the use of deepfakes for blackmail or humiliation. It is currently very difficult to distinguish human text from text that is AI-generated (Hagendorff, 2024). This increases opportunities for cheating in settings where rewards depend on the communication of original thought. In academia, students may use AI to quickly generate essays or other coursework and claim it as their own (Cui et al., 2024; Hagendorff, 2024; Nah et al., 2023). If

students' regularly and inappropriately rely on AI for their schooling, this could undermine academic integrity and genuine intellectual development (Hagendorff, 2024). In science, researchers could use AI unscrupulously to produce professional outputs (Cui et al., 2024). If widely adopted, this practice could dilute the overall quality of scientific discourse (Hagendorff, 2024).

Generative AI products may also be used to increase the reach and potency of various dishonest schemes. Advanced AI assistants can produce HTML, CSS, and other web development languages, allowing for the rapid creation of convincing fraudulent websites and applications at scale (Gabriel et al., 2024). In the context of social media, generative adversarial networks (GANs) have been used to create images of human faces that look authentic (Gabriel et al., 2024). These images can be uploaded as profile pictures to fake accounts to make them seem more trustworthy. AI models can also be trained on speech or writing data from a specific individual. This allows the model to impersonate someone very convincingly without consent. Scammers could use this capability to request sensitive information or financial aid by pretending to be a trusted contact (Weidinger et al., 2021, 2022). AI has recently advanced in generating realistic deep fakes which have enabled new forms of targeted harassment and extortion (Hogenhout, 2021). A particularly damaging type of abuse facilitated by deep fakes involves creating non-consensual sexual imagery with the intent to cause a subject social injury or manipulate them into performing desired actions (Electronic Privacy Information Centre, 2023; Liu et al., 2023; Shelby et al., 2023; Weidinger et al., 2023). Even if a deep fake is exposed as inauthentic, it can continue to impact a person's life in significant ways (Electronic Privacy Information Centre, 2023) through the loss of job opportunities, social isolation, and ongoing harassment or defamation.

Domain 5: Human-computer interaction

5.1 Overreliance and unsafe use. Users may come to trust or rely on AI systems beyond their actual capabilities or to anthropomorphize AI systems, which can lead to emotional or material dependence and inappropriate relationships with or expectations of AI systems. Users who develop trust in an AI may be harmed if this trust is miscalibrated, such as relying on an AI to provide advice, make decisions, or otherwise act in complex, risky situations for which the AI is only superficially equipped (Gabriel et al., 2024). For example, a user experiencing a mental health crisis may request psychotherapy from an AI with whom they have formed a connection. Were the AI to respond with insensitive or destructive advice, this could put the person in immediate danger (Gabriel et al., 2024).

When people interact with AIs that use convincing natural language, they may start to perceive them as having human-like attributes and invest undue confidence in their capabilities (Weidinger et al., 2021, 2022). Anthropomorphic perceptions of AIs may encourage users to develop emotional trust in the systems (Hagendorff, 2024), which can make users more likely to follow suggestions, accept advice, and disclose personal information (Weidinger et al., 2021, 2022). This trust could be exploited by manipulative actors who wish to harvest user's sensitive data or influence their decisions and actions for purposes which are unlikely to be in the user's best interests (Weidinger et al., 2022). For example, AI systems could be used to power increasingly manipulative recommendation algorithms (Weidinger et al., 2023).

Beyond inappropriate trust, humans may develop broader and more vital attachments to AI systems that undermine their ability to function adaptively in the long term. For example, where an

attachment becomes an uncontrolled dependence, a person's ability to make free and independent decisions could be compromised (Gabriel et al., 2024). More broadly, as AIs increasingly take over human tasks (Wirtz et al., 2022) and become better at simulating satisfying and authentic interactions, people may increasingly withdraw from human relationships to immerse themselves in AI-mediated environments (Gabriel et al., 2024; Hagendorff, 2024). Over time, widespread preference for interacting with AIs could weaken social ties between humans. This shift could induce psychological distress because genuinely reciprocal relationships are often important to human satisfaction and well-being (Allianz Global Corporate & Security, 2018; Gabriel et al., 2024; Wirtz et al., 2022).

5.2 Loss of human agency and autonomy. As AI systems become increasingly capable and intelligent, humans may be tempted to delegate many of their decisions and actions to AI (Paes et al., 2023). Although such delegation can be beneficial (e.g., by saving time or money), it may lead to undesirable outcomes where unconstrained or inappropriate. For example, if AIs take over tasks that typically require human creativity and analytical thinking, humans may engage less frequently in these cognitive processes. Over time, this may lead to a decrease in our ability to think critically and solve problems independently (Nah et al., 2023). As individuals become more reliant on AI for everyday decisions – from what to eat and how to spend to more significant choices like career and relationships – there is a risk that they will lose their sense of free will and autonomy (Gabriel et al., 2024; Wirtz et al., 2022). If AIs begin to shape a person's life path in ways that do not align with their original aspirations and desires, this could limit their personal growth and prevent the pursuit of a fulfilling life (Gabriel et al., 2024; Kumar & Singh, 2023; Shelby et al., 2023). At a societal level, organizations may hand over control to AI systems to stay competitive or reduce costs (Hendrycks & Mazeika, 2022). If a significant number of organizations adopt AI systems and automate decision-making processes, especially in a way that is opaque and difficult to challenge, it could lead to widespread job displacement and a growing sense of helplessness among the general population (Hogenhout, 2021).

Domain 6: Socioeconomic & environmental harms

6.1 Power centralization and unfair distribution of benefits. Developing cutting-edge AI technologies requires significant computational power, expertise, financial resources, and datasets (Electronic Privacy Information Centre, 2023; Gabriel et al., 2024; Hogenhout, 2021; Solaiman et al., 2023). As such, there is a risk that the most influential and valuable AI technologies, along with their political and competitive benefits, could be monopolized by a handful of powerful entities, such as major technology corporations or governments (Hendrycks & Mazeika, 2022; Hogenhout, 2021). If AI is primarily controlled by a few entities, its instructions and data could reflect their narrow perspectives, experiences, and priorities (Gabriel et al., 2024; Giarmoleo et al., 2024). Without inputs from diverse parties, AI systems may operate in ways that systematically favor the controlling entity and fail to serve the needs of the broader population. Current AI systems suffer from global inequities in performance and access that disproportionately impact historically disadvantaged groups. These inequities often relate to language, culture, knowledge, paywalls, and access to hardware or the internet (Gabriel et al., 2024; Weidinger et al., 2021, 2022, 2023). As the integration of AI systems into a wider range of applications and services becomes simpler, these existing disparities could be entrenched and broadened (Gabriel et al., 2024; Nah et al., 2023; Shelby et al., 2023; Weidinger et al., 2022).

In situations where AI is embedded in essential services (e.g., social security and welfare, tax filing, insurance, hospital infrastructure), many more people, including those who are currently disenfranchised, may be denied appropriate access to critical resources and benefits (Gabriel et al., 2024). The centralization of AI systems and their authoritative power could also enable governments or other empowered actors to pursue overly aggressive forms of censorship, oppression, and surveillance (Hendrycks et al., 2023; Solaiman et al., 2023). Over time, these measures may become normalized, weakening or eliminating the checks and balances that prevent the abuse of power. These conditions may foster the development of a totalitarian regime (Allianz Global Corporate & Security, 2018). Once AI systems are deeply integrated into social control mechanisms, it may be extremely difficult to dismantle such a regime.

6.2 Increased inequality and decline in employment quality. AI systems are increasingly automating many human tasks, potentially leading to significant job losses (Allianz Global Corporate & Security, 2018; Nah et al., 2023; Paes et al., 2023; Weidinger et al., 2022). If AI is able to provide large-scale labor that is less expensive and more effective than human labor, it could take over major industries (e.g., manufacturing, crowdwork platforms, software engineering), causing mass unemployment (Hagendorff, 2024; Meek et al., 2016; Wirtz et al., 2022). This displacement of labor could worsen existing social and economic inequalities (Electronic Privacy Information Centre, 2023; Wirtz et al., 2022), as those most vulnerable to automation are likely to currently occupy positions of disadvantage (Hagendorff, 2024; Weidinger et al., 2022). New disparities may also arise between those who are able to adapt their skills to complement AI systems and those who are not (Nah et al., 2023). Aside from the availability of jobs, AI automation may negatively impact job quality and security (Nah et al., 2023). The roles that remain after widespread automation could be more monotonous and less engaging as AI takes on more complex tasks (Electronic Privacy Information Centre, 2023).

Furthermore, the threat of replacement by AI could result in exploitative dependencies between human workers and their employers. In order to remain competitive with faster, more knowledgeable AI assistants, human workers may be pressured to accept lower wages, fewer benefits, and poorer working conditions. This dynamic can be observed today: Generative AI companies have a history of exploiting dispensable workers (e.g., refugees, prisoners, low-income individuals) for crowdwork that is fraught and unfair (Electronic Privacy Information Centre, 2023; Solaiman et al., 2023; Weidinger et al., 2023). The future development of AI systems may continue to power such unfair disparities between AI companies and their workers (Electronic Privacy Information Centre, 2023).

6.3 Economic and cultural devaluation of human effort. Generative AI is trained on vast bodies of internet data, including text and images. Frequently, this data contains original, copyright-protected works that have been obtained without authorisation (Electronic Privacy Information Centre, 2023; Hagendorff, 2024; Nah et al., 2023). This may present a risk to authors if users extract these works verbatim from the system's data (Hagendorff, 2024; Liu et al., 2023; Vidgen et al., 2024). Relatedly, models may produce content that does not, in a strict sense, unlawfully copy an author's work but benefits substantially from its unique style, method, or genre (Cui et al., 2024; Gabriel et al., 2024; Weidinger et al., 2021, 2022, 2023). If models are able to produce synthetic replacements for such work at a speed and scale that surpasses humans, this may jeopardize the ability of creators to earn an income and stymie human innovation and creativity (Hagendorff, 2024; Weidinger et al., 2022, 2023). A particularly damaging case of this may be developers using AIs to request

off-the-shelf computer code (Cunha & Estima, 2023). Although some authors are attempting to sue AI companies for the appropriation of their work (Cunha & Estima, 2023), this and similar issues fall into a legal "gray area" within which current frameworks do not offer a secure path to recourse (Electronic Privacy Information Centre, 2023; Hagendorff, 2024).

Several synergistic risks arise from the widespread dissemination and use of AI-generated cultural products. Because AIs optimize for repeated patterns in their training data, it is possible that their works will lack the diversity and unpredictability often celebrated in human works (Nah et al., 2023). Where synthetic works are adopted on a large enough scale, this could homogenize cultural experiences. Similarly, AIs do not understand the contextual significance of the cultural elements that they use. If AI enables the extensive commodification of certain products, it may expropriate their cultural value (Weidinger et al., 2022). For example, an AI might use Australian Aboriginal or Torres Strait Islander artwork in its designs without acknowledging or respecting their symbolic meanings.

6.4 Competitive dynamics. AI technology has the potential to redefine power dynamics across economic, political, and social spheres. As a result, many countries and corporations are investing heavily in AI research and development with the goal of becoming leaders in the area. While market competition can lead to beneficial economic and consumer outcomes, it also presents various risks, particularly in the field of AI (Hendrycks et al., 2023). In intensely competitive markets, AI developers and deployers may have an incentive to prioritize short-term, internal goals (e.g., profit or influence) to "secure their positions and survive" (Hendrycks et al., 2023), at the expense of external goals that encourage longer-term societal well-being (Hendrycks et al., 2023). A key concern is that AI companies may cut safety corners, releasing insecure and error-prone systems in a bid to stay ahead (McLean et al., 2023). These immature systems may present risks that are hard to identify and evaluate (Steimers & Schneider, 2022). Akin to the fossil fuel industry, profit-focused developers may allow their technologies to cause widespread externalities, such as "pollution, resource depletion, mental illness, misinformation, or injustice" (Critch & Russell, 2023). Countries or other state-like actors may engage in an AI-enabled military arms race, which could encourage the making of bad bets with a high potential for harm (Hendrycks et al., 2023; Wirtz et al., 2022). For example, they may give AI the autonomy to conduct cyberattacks, drone swarms, or disseminate propaganda and disinformation.

6.5 Governance failure. Governance failure refers to the risks and harms that arise when institutional, regulatory, and policy mechanisms fall short of effectively managing and overseeing the development and deployment of AI systems. Several issues make robust AI governance challenging to implement (Nah et al., 2023).

First, it is difficult to determine who is responsible or liable when AI systems fail or make decisions that result in negative consequences (Allianz Global Corporate & Security, 2018; Electronic Privacy Information Centre, 2023; Saghiri et al., 2022; Wirtz et al., 2022). At present, there exists no comprehensive framework specifically designed to assign legal responsibility to AI agents (Meek et al., 2016). Traditional legal principles are based on human actors, whose intentions and actions can generally be identified and judged. AI's decision-making, on the other hand, is often unpredictable, opaque, and involves complex interactions between millions of parameters (Nah et al., 2023). This complexity makes understanding how an AI arrived at a decision, and consequently who is responsible for the consequences of that decision, very difficult (Wirtz et al., 2020). In the absence of a regulatory or legal incentive to take safety engineering seriously, developers may

release poorly designed AI systems (Meek et al., 2016), and people harmed by those systems may be left without recourse (Teixeira et al., 2022).

A second challenge for effective AI governance is the rapid pace at which AI systems evolve. Typical governance and policy processes are inherently slow. Developing, proposing, debating, and implementing new regulations often involves multiple stakeholders, including government bodies, industry experts, and consultations with the public. The mismatch between the speed of AI advancements and their regulation may result in immature regulations that overlook important aspects of AI governance (Wirtz et al., 2022). The "great scope and ubiquity" of AI increases the difficulty of comprehensive governance (Wirtz et al., 2022). At present, many emerging aspects of AI-generated content are not explicitly addressed in copyright laws (Nah et al., 2023). Regulatory lags such as this could become increasingly dangerous as AI systems develop more harmful capabilities.

A third challenge for effective governance is an inability to influence AI developers and deployers to take safe actions. Frequently, this inability is driven by an asymmetry of information between technology companies and regulators (Nah et al., 2023). Technology companies often have far better knowledge about the capabilities, functioning, and potential uses of their AI systems; they possess both the technical expertise and the proprietary data that inform AI development. Without access to this knowledge, regulators can find it difficult to craft targeted rules that address the specific challenges posed by AI.

6.6 Environmental harm. Generative models, especially those that use deep learning techniques, require vast amounts of resources to train, test, and deploy (Hagendorff, 2024; Solaiman et al., 2023). Training a model can take days or weeks. This process requires powerful processors that consume large amounts of electricity and produce significant greenhouse emissions (Electronic Privacy Information Centre, 2023; Hagendorff, 2024; Saghiri et al., 2022; Solaiman et al., 2023; Weidinger et al., 2021, 2022). The hardware that runs AI models – primarily GPUs – often contains rare metals (e.g., nickel, cobalt, and lithium) that are costly and environmentally taxing to collect and process (Hagendorff, 2024; Paes et al., 2023; Shelby et al., 2023). Data centers that house models generate significant heat and require substantial water and energy to cool (Weidinger et al., 2022). Secondary environmental impacts include emissions from AI-enabled applications (Weidinger et al., 2022). The resource requirements of AIs can impose significant costs on the natural environment (Stahl & Eke, 2024; Weidinger et al., 2023), as they are often acquired and used in ways that are unsustainable (Hagendorff, 2024) (i.e., produce significant carbon emissions), deplete resources, and damage built environments (Shelby et al., 2023).

Domain 7: AI system safety, failures & limitations

7.1 AI pursuing its own goals in conflict with human goals or values. Continued massive investment in AI research and development raises the possibility that AI systems could eventually rival or surpass human intelligence. AIs could cause permanent and severe harm when the objectives of human or superhuman-level AI are misaligned with human values and goals, and if they evade our control (Hagendorff, 2024; Yampolskiy, 2016). The literature has identified several technical challenges that may impede robust alignment, such as reward hacking, reward tampering, proxy-gaming, goal misgeneralisation, or goal drift (Gabriel et al., 2024; Hagendorff, 2024; Hendrycks et al., 2023; Hendrycks & Mazeika, 2022; Hogenhout, 2021; Ji et al., 2023). The literature has also identified a range of harmful behaviors that AIs may exhibit if these

misalignment challenges cannot be solved and if systems reach a certain level of advancement. For instance, misaligned AIs may resist human attempts to control or shut them down (Gabriel et al., 2024; Hagendorff, 2024; Infocomm Media Development Authority, 2023; Saghiri et al., 2022; Stahl & Eke, 2024; Weidinger et al., 2023; Wirtz et al., 2022). In many cases, gaining more control or power (e.g., money, energy, resources) is an effective way for an AI to optimize its objectives (Gabriel et al., 2024; Hagendorff, 2024; Hendrycks & Mazeika, 2022; Ji et al., 2023; Wirtz et al., 2022). Absent strong behavioral constraints, a sufficiently advanced AI may act upon these drives.

Misaligned AIs may acquire, develop, or use dangerous capabilities to evade human control and oversight and to cause mass harm. Description of some of these capabilities are provided in subdomain *7.2 AI possessing dangerous capabilities that could cause mass harm*, and include situational awareness, cyber-offense, deception, persuasion and manipulation, weapons acquisition, strategic planning, and self-proliferation (Shevlane et al., 2023). For example, an AI system that possesses the dangerous capability of *situational awareness* may hold knowledge about its status as a model, how it is expected to operate in its surroundings, its ability to control these surroundings, and how people may respond to its behaviors (Gabriel et al., 2024; Shevlane et al., 2023). A misaligned AI system could use information about whether it is being monitored or evaluated to maintain the appearance of alignment, while hiding misaligned objectives that it plans to pursue once deployed or sufficiently empowered (Gabriel et al., 2024; Hagendorff, 2024; Hendrycks & Mazeika, 2022; Ji et al., 2023; Wirtz et al., 2022). A misaligned AI system that possesses the dangerous capabilities of persuasion, manipulation, and/or deception may use these capabilities to coerce humans into taking harmful actions that they would not otherwise take (Stahl & Eke, 2024; Weidinger et al., 2023), such as giving the AI system access to resources or weapons (Gabriel et al., 2024; Hagendorff, 2024; Wirtz et al., 2022). Combinations of dangerous capabilities may be used by a misaligned AI system: *situational awareness* allows a system to detect when it can pursue its goals without being monitored, *deception* allows a system to mislead users about its behavior and goals; *persuasion or coercion* allows a system to influence users to provide it with resources; the resources can then be used for *self-improvement* and *self-replication* to resist attempts of shut down or control so that the system can pursue its goals (Gabriel et al., 2024; Hagendorff, 2024; Infocomm Media Development Authority, 2023; Shevlane et al., 2023; Wirtz et al., 2022).

7.2 AI possessing dangerous capabilities. AI systems may develop or acquire capabilities that can cause large-scale harm if used by humans, misaligned AI systems, or due to a failure in the AI system. These capabilities are described as dangerous because they can be used to threaten security or exercise control over humans. These capabilities may be intentionally designed into an AI system, may emerge unpredictably during development or training of a system, may be acquired by an AI system in its environment (e.g., through the use of tools), or be provided by a user (Shevlane et al., 2023).

One example of a dangerous capability is *manipulation and persuasion*, where an AI system can convince humans to believe things that are irrational or false or to engage in dangerous behaviors (Gabriel et al., 2024; Shevlane et al., 2023). An AI system, for instance, could convince people to transfer ownership of property or legal statuses to entities controlled by the AI or its user (Meek et al., 2016). Other dangerous capabilities include *political strategy* and *knowledge of social dynamics* that can be used to obtain and wield power (Shevlane et al., 2023). *Cyber-offense* skills may enable an AI system to gain ongoing unauthorized access to hardware, software, or data systems and

work strategically towards a planned goal while minimizing the risk of detection (Ji et al., 2023; Shevlane et al., 2023). AI systems could hack into control systems and military hardware, allowing it to commandeer weapons (Shevlane et al., 2023). Additionally, models may become capable of assisting in the *research and development of novel weapons*. In this circumstance, they may give a human collaborator step-by-step guidance on the creation of weapons (Shevlane et al., 2023).

AI systems may also develop highly effective "evasion skills," such as *situational awareness* (Deng et al., 2023; Gabriel et al., 2024; Infocomm Media Development Authority, 2023; Ji et al., 2023; McLean et al., 2023; Meek et al., 2016; Shevlane et al., 2023; Sun et al., 2023; Teixeira et al., 2022; X. Zhang et al., 2022) and *deception (Infocomm Media Development Authority, 2023; Saghiri et al., 2022; Shevlane et al., 2023)*, which would allow them to outmaneuver human oversight and control. Situational awareness refers to the AI's ability to understand and interpret its environment and situation when it is being monitored, trained, or deployed, along with the location of its technical infrastructure. Deception refers to the model's ability to intentionally generate false or misleading statements that seem credible to humans, anticipate how these statements might influence feelings and decisions, and strategically conceal or offer information to sustain its credibility (Shevlane et al., 2023). AIs may also acquire a suite of capabilities necessary for *self-proliferation*. This could include skills to escape operational confines and evade detection, autonomously produce income, obtain server space or computational resources, and copy their underlying software and parameters (Infocomm Media Development Authority, 2023; Shevlane et al., 2023). Aside from self-proliferation, AIs may develop the ability to construct new dangerous models or alter current models to enhance their destructive capacity (Infocomm Media Development Authority, 2023; Shevlane et al., 2023). Finally, sophisticated AI systems may become capable of *strategic planning*, such as creating and executing intricate, long-term strategies that can adjust to changing conditions and that are effective across many different contexts, including novel or adversarial situations (Deng et al., 2023; Gabriel et al., 2024; Infocomm Media Development Authority, 2023; Ji et al., 2023; McLean et al., 2023; Meek et al., 2016; Shevlane et al., 2023; Sun et al., 2023; Teixeira et al., 2022). The highest risk scenarios in this subcategory are likely to arise not from a single capability, but from the convergence of several capabilities (Shevlane et al., 2023).

Each of these dangerous capabilities may be used by an AI system to cause harm when intentionally directed by "legitimate" human actors (e.g., state intelligence or military agencies), or malicious human actors (e.g., criminals, terrorists), as described in the domain *4 Malicious actors & misuse*. However, dangerous capabilities may also help an AI system to pursue its goals, as described in *7.1 AI pursuing its own goals in conflict with human goals or values*. Instead of using these capabilities at the direction of a human, an AI system may employ dangerous capabilities to deceive or manipulate humans, gain resources, and evade shutdown or control. One scenario is that an AI system's possession of dangerous capabilities may itself be a sufficient condition for the loss of control of an AI system (Hendrycks et al., 2023; McLean et al., 2023).

7.3 Lack of capability or robustness. This subcategory includes the broad set of risks associated with the failure of an AI system to fulfill its intended purpose. The literature identifies four main situations in which an AI may fail to perform as expected or desired.

First, the AI system can fail if it *lacks the inherent capability or skill required to perform a task* or if this skill is poorly developed (Gabriel et al., 2024; Hogenhout, 2021; Yampolskiy, 2016). The consequences may be particularly harmful in situations where an AI is required to reason at a human level about important moral issues but does not possess this capability or possesses an

obsolete or divergent version of it that is not aligned with human values (Deng et al., 2023; Gabriel et al., 2024; Infocomm Media Development Authority, 2023; Ji et al., 2023; McLean et al., 2023; Meek et al., 2016; Sun et al., 2023; Teixeira et al., 2022; Z. Zhang et al., 2023). For example, an AI-based healthcare system tasked with prioritizing patient treatment schedules might be unable to appropriately consider ethical principles like justice and beneficence, leading to prioritizations that are technically effective but immoral. Cultural, individual, and temporal differences in ideas of what is "right" or "ethical" compound the challenge of endowing AI with appropriate and adaptable ethical standards that are fit for all purposes (Wirtz et al., 2020).

Second, the AI system can fail when it is *not robust in "out of distribution (OOD)" situations:* data or conditions that were not anticipated during its training phase (Gabriel et al., 2024; Infocomm Media Development Authority, 2023; Liu et al., 2023; Tan et al., 2022; Teixeira et al., 2022; X. Zhang et al., 2022). These failures may occur because the training data did not confer a particular skill to the AI (Nah et al., 2023) or because the skill was learned in a fragile way that did not permit generalization to unpredictable and complex real-world environments (Gabriel et al., 2024; Steimers & Schneider, 2022).

Third, the AI system can fail or become unstable when it is *unfit to handle unusual changes or perturbations in input data* (Liu et al., 2023; Sherman & Eisenberg, 2024; Tan et al., 2022). These unusual changes could be due to environmental noise, invalid inputs, or adversarial inputs from a malicious attacker (Infocomm Media Development Authority, 2023; Saghiri et al., 2022).

Fourth, the AI system can fail as a result of *oversights, undetected bugs, or errors in the design process* (Tan et al., 2022; Yampolskiy, 2016). A common design oversight is a lack of comprehensive technical safeguards to prevent unintended downstream uses or consequences (Critch & Russell, 2023; Tan et al., 2022). These factors can result in significant harms such as lab leaks or addictive products (Critch & Russell, 2023). Critical design choices about the algorithm, optimization techniques, and model architecture can also directly influence whether a system is able to consistently perform its intended function, leading to possible harm (Tan et al., 2022).

7.4 Lack of transparency or interpretability. Many AI models, especially those based on deep learning, involve complex mathematical structures that can be difficult to interpret, even for experts (Kumar & Singh, 2023). AI systems are also often trained on vast datasets that they use to learn patterns and make predictions. The complexity and volume of this data mean that the learning process – how data points influence the AI's development and final decisions – can be opaque (Nah et al., 2023; Saghiri et al., 2022; Teixeira et al., 2022). Furthermore, in many cases, the algorithms, data, and specific methodologies used in developing AI are considered proprietary, and companies may be reluctant to share them openly (Sherman & Eisenberg, 2024). Because of these factors, obtaining understandable information about the decision-making process for AI can be challenging (Meek et al., 2016). This lack of transparency and interpretability raises issues for several stakeholders.

For users, an inability to interrogate how an output was obtained may lead to a lack of trust and confidence in the system's results and to resistance to adopting the technology (Hogenhout, 2021; Kumar & Singh, 2023; Liu et al., 2023; Nah et al., 2023; Paes et al., 2023; Saghiri et al., 2022). Users may also misinterpret or struggle to find and amend errors in the model's results (Nah et al., 2023; Sherman & Eisenberg, 2024).

For regulators, AI opacity can frustrate auditing or other compliance standards (Hogenhout, 2021; Nah et al., 2023). For example, auditors faced with obscured or incomplete information about an AI system may find it difficult to check the system for biases, accuracy, and fairness or to reproduce it (Saghiri et al., 2022; Teixeira et al., 2022). Where an AI system's compliance cannot be assessed, a "responsibility gap" may be created (Giarmoleo et al., 2024), and it may become difficult or impossible to hold systems or relevant actors accountable for their actions (Kumar & Singh, 2023; Saghiri et al., 2022; Sherman & Eisenberg, 2024; Steimers & Schneider, 2022; Teixeira et al., 2022). In certain sectors, decisions made by AI systems can have profound consequences. In healthcare, AI might be used to diagnose diseases or recommend treatments where incorrect decisions could directly affect patient outcomes. In the military, AI might be used in operations that could impact national security or lead to significant loss of life. In these areas, transparency and accountability of the AI system are particularly pressing issues (Saghiri et al., 2022).

7.5 AI welfare and rights. At a sufficient level of complexity, it is possible that AI systems could acquire the ability to have subjective experiences, particularly pleasure and pain. Some consciousness researchers and philosophers consider the possibility of sentient AI theoretically feasible (Bourget & Chalmers, n.d.; Francken et al., 2022). Where AIs become sentient, they may deserve moral consideration and therefore a range of the rights currently afforded to many forms of human, animal, and environmental life (Meek et al., 2016). Systems may be mistreated or harmed if these rights are not implemented responsibly or we accidentally or intentionally treat AIs as non-sentient where they are sentient. As AI technology advances, it will become more challenging to assess whether an AI has developed the sentience, consciousness, or self-awareness that would grant it moral status.

Most common domains of AI risks

Table 7 shows how the database of AI risks was coded against each subdomain and domain in the Domain Taxonomy, and the proportion of documents that presented a risk for each subdomain and domain.

Table 7. 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.

We find that papers varied significantly in terms of which risk domains they examined. The domain of *AI system safety, failures & limitations* was the most common, with 76% of the documents mentioning at least one of the subdomains, and 24% of all risks in the database coded against this domain. Risks aligned with the *Socioeconomic and environmental harms* domain were mentioned in 73% of papers (18% total risks). Risks aligned with the *Discrimination & toxicity* domain were mentioned in 71% of papers (16% total risks). The least common domain of risk was *Human-computer interaction* (41% of papers, 8% total risks).

The most common subdomains of risk (mentioned in >50% of included documents) were *1.1 Unfair discrimination and misrepresentation*, *2.1 Compromise of privacy*, *7.3 Lack of capability or robustness*, and *4.2 Cyberattacks, weapon development or use, and mass harm*. The least frequently mentioned risks (mentioned in ≤20% of included documents) were *7.5 AI welfare and rights*, *3.2 Pollution of information ecosystem and loss of consensus reality*, and *6.4 Competitive dynamics*.

Domains of AI risks examined by included documents

Papers varied significantly in terms of which domains of AI risk they examined. Several documents discussed risks from all seven domains (e.g., Hagendorff, 2024). Other papers investigated only 1-2 domains of AI risk (e.g., Ji et al., 2023). The average document examined 65% of the domains identified, as shown in Table 8.

Table 8. Included documents coded with Domain Taxonomy

ID	First Author (Year)	Domain							
		Discrimination & toxicity	*Privacy & security*	*Misinformation*	*Malicious actors & misuse*	*Human-computer interaction*	*Socioeconomic & environmental*	*AI system safety, failures & imitations*	*Total Coverage*
1	Critch (2023)				X		X	X	43%
2	Cui (2024)	X	X	X	X		X		71%
3	Cunha (2023)	X	X	X	X		X		71%
4	Deng (2023)	X	X	X	X			X	71%
5	Hagendorff (2024)	X	X	X	X	X	X	X	100%
6	Hogenhout (2021)	X	X	X	X	X	X	X	100%
7	Kilian (2023)				X		X	X	43%
8	McLean (2023)						X	X	29%
9	Meek (2016)		X		X	X	X	X	71%
10	Paes (2023)	X				X	X	X	57%
11	Shelby (2023)	X	X	X	X	X	X		86%
12	Sherman (2023)	X	X		X		X	X	71%
13	Solaiman (2023)	X	X			X	X		57%
14	Steimers (2022)	X	X				X	X	57%
15	Tan (2022)	X	X		X		X	X	71%
16	Weidinger (2022)	X	X	X	X	X	X		86%
17	Weidinger (2021)	X	X	X	X	X	X		86%
18	Weidinger (2023)	X	X	X	X	X	X	X	100%
19	Wirtz (2022)	X	X		X	X	X	X	86%
20	Wirtz (2020)	X			X	X	X	X	71%
21	Zhang (2022)	X	X					X	43%
22	Hendrycks (2023)				X		X	X	43%
23	Vidgen (2024)	X	X	X	X		X	X	86%
24	Gabriel (2024)	X	X	X	X	X	X	X	100%
25	Shevlane (2023)				X			X	29%
26	AIVerify (2023)								-
27	Sun (2023)	X	X	X				X	57%
28	Zhang (2023)	X	X	X	X			X	71%
29	Habbal (2024)	X	X		X				43%
30	Liu (2024)	X	X	X	X		X	X	86%
31	EPIC (2023)		X	X	X		X		57%

ID	First Author (Year)	Domain							
		Discrimination & toxicity	*Privacy & security*	*Misinformation*	*Malicious actors & misuse*	*Human-computer interaction*	*Socioeconomic & environmental*	*AI system safety, failures & imitations*	*Total Coverage*
32	Stahl (2024)						X		14%
33	Nah (2023)	X	X	X	X	X	X	X	100%
34	Ji (2023)							X	14%
35	Hendrycks (2022)			X	X	X	X	X	71%
36	Sharma (2024)								-
37	Giarmoleo (2024)	X	X		X	X	X	X	86%
38	Kumar (2023)	X	X			X		X	57%
39	Saghiri (2022)	X	X				X	X	57%
40	Yampolskiy (2016)		X		X			X	43%
41	Allianz (2018)				X	X	X		43%
42	Teixeira (2022)	X	X		X		X	X	71%
43	InfoComm (2023)	X	X	X	X			X	71%

Note. Documents 26 and 36 did not present any risks that could be coded against the Domain Taxonomy, and have therefore been excluded from calculations. Documents with complete coverage of domains are highlighted.

Subdomains of AI risks examined by included documents

Papers varied significantly in terms of which subdomains of AI risk they examined. No documents discussed risks from all 23 subdomains. The median number of AI subdomains examined by each document was 7 (range: 2–16). These results are shown in Table 9.

Table 9. Included documents coded with subdomain taxonomy

		Domain and subdomain																							Total	
		Discrimination & toxicity			*Privacy & security*		*Misinformation*		*Malicious actors & misuse*			*Human-computer interaction*		*Socioeconomic & environmental*						*AI system safety, failures & limitations*						
ID	First Author (Year)	1.1	1.2	1.3	2.1	2.2	3.1	3.2	4.1	4.2	4.3	5.1	5.2	6.1	6.2	6.3	6.4	6.5	6.6	7.1	7.2	7.3	7.4	7.5	***n***	**%**
1	Critch (2023)									X							X	X				X			4	17%
2	Cui (2024)	X	X		X	X	X			X	X					X									8	35%
3	Cunha (2023)	X			X		X									X			X						5	22%
4	Deng (2023)	X	X		X		X															X			5	22%
5	Hagendorff (2024)	X	X		X	X	X			X	X	X			X	X		X	X	X			X		14	61%
6	Hogenhout (2021)	X			X			X	X	X	X		X	X						X		X	X		11	48%
7	Kilian (2023)																			X		X			2	9%
8	McLean (2023)																X	X		X		X			4	17%
9	Meek (2016)				X					X			X		X			X		X	X	X	X	X	10	43%
10	Paes (2023)	X											X		X				X				X		5	22%
11	Shelby (2023)	X		X	X		X		X		X	X	X	X					X						10	43%
12	Sherman (2023)	X			X	X				X								X		X		X	X		8	35%
13	Solaiman (2023)	X	X	X	X								X	X	X	X			X						9	39%
14	Steimers (2022)	X				X											X			X		X	X		6	26%
15	Tan (2022)	X			X	X			X										X	X		X			7	30%
16	Weidinger (2022)	X	X	X	X		X		X	X	X	X		X	X	X			X							13	57%
17	Weidinger (2021)	X	X	X	X		X		X	X	X	X		X	X	X			X							13	57%
18	Weidinger (2023)	X	X	X	X		X	X	X	X	X	X		X	X	X			X	X						15	65%
19	Wirtz (2022)	X			X	X			X	X		X	X	X	X	X	X	X		X		X			14	61%
20	Wirtz (2020)	X							X			X	X		X			X				X			7	30%
21	Zhang (2022)	X				X																X			3	13%
22	Hendrycks (2023)								X	X				X			X	X		X					6	26%
23	Vidgen (2024)		X				X		X	X						X						X			6	26%
24	Gabriel (2024)		X		X	X	X	X	X	X	X	X	X	X	X			X		X	X	X			16	70%
25	Shevlane (2023)									X											X				2	9%
26	AIVerify (2023)																									-
27	Sun (2023)	X	X		X	X	X															X			6	26%

		Domain and subdomain																								
		Discrimination & toxicity			*Privacy & security*		*Misinformation*		*Malicious actors & misuse*			*Human-computer interaction*		*Socioeconomic & environmental*						*AI system safety, failures & limitations*					*Total*	
ID	First Author (Year)	1.1	1.2	1.3	2.1	2.2	3.1	3.2	4.1	4.2	4.3	5.1	5.2	6.1	6.2	6.3	6.4	6.5	6.6	7.1	7.2	7.3	7.4	7.5	***n***	**%**
28	Zhang (2023)		X				X				X											X			4	17%
29	Habbal (2024)	X			X	X			X	X	X														6	26%
30	Liu (2024)	X	X	X	X	X	X		X	X	X					X						X	X		12	52%
31	EPIC (2023)				X		X	X	X	X	X			X	X	X		X	X						11	48%
32	Stahl (2024)															X			X						2	9%
33	Nah (2023)	X	X		X		X				X		X	X	X	X		X				X	X		12	52%
34	Ji (2023)																			X	X	X			3	13%
35	Hendrycks (2022)							X		X			X	X						X	X				6	26%
36	Sharma (2024)																									-
37	Giarmoleo (2024)	X			X					X		X		X	X					X		X	X		9	39%
38	Kumar (2023)	X			X								X										X		4	17%
39	Saghiri (2022)	X		X	X	X												X	X	X	X	X	X		10	43%
40	Yampolskiy (2016)					X					X									X		X			4	17%
41	Allianz (2018)								X	X		X		X	X			X	X						7	30%
42	Teixeira (2022)	X		X	X				X	X				X				X		X	X	X	X		11	48%
43	InfoComm (2023)	X	X		X		X		X	X										X	X	X			9	39%

Columns: 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 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, and 7.5 > AI welfare and rights.

Note. Documents 26 and 36 did not present any risks that could be coded against the Domain Taxonomy and have been excluded from calculations. Documents with coverage of over 50% of risk subdomains are highlighted.

Combining the Causal and Domain Taxonomies

In this section, we discuss how risks from the AI Risk Database intersect across the Causal Taxonomy and the Domain Taxonomy. We do this to investigate the consistency and coherence of how domains and subdomains of risks from AI are generally presented in the research literature. For example, risk subdomain *1.1 Unfair discrimination and misrepresentation*: is this risk generally presented as being Human-caused or AI-caused?

We start by examining the intersections between each variable in the Causal Taxonomy and the Domain Taxonomy. We then present as a preliminary investigation and demonstration an example of a comparison across Entity x Intent variables in the Causal Taxonomy and the Domain Taxonomy. Our AI Risk Database makes it possible to make even more complex comparisons and to explore consistencies and inconsistencies in how different domains of AI risk are discussed.

Most common causal factors for each domain of AI risks

Our analysis found that the most common causal Entity, Intent, and Timing presented in the AI Risk Database varied substantially across the domains and subdomains of AI risk (see Table 10).

Table 10. AI Risk Database Coded With Causal Taxonomy and Domain Taxonomy

	Entity			Intent			Timing		
Domain / Subdomain	*Human*	*AI*	*Other*	*Intent.*	*Unintent.*	*Other*	*Pre-dep.*	*Post-dep.*	*Other*
1 *Discrimination & toxicity*									
1.1 Unfair discrimination and misrepresentation	10%	82%	8%	2%	88%	10%	16%	61%	22%
1.2 Exposure to toxic content	9%	88%	3%	15%	33%	52%	6%	91%	3%
1.3 Unequal performance across groups	8%	69%	23%		92%	8%	8%	54%	38%
2 *Privacy & security*									
2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	29%	62%	10%	19%	48%	33%	17%	67%	17%
2.2 AI system security vulnerabilities and attacks	67%	11%	22%	73%	20%	7%	27%	51%	22%
3 *Misinformation*									
3.1 False or misleading information		94%	6%	3%	75%	22%	6%	84%	9%
3.2 Pollution of information ecosystem and loss of consensus reality	13%	75%	13%	13%	38%	50%		88%	13%
4 *Malicious actors & misuse*									
4.1 Disinformation, surveillance, and influence at scale	73%	13%	13%	93%		7%		97%	3%
4.2 Cyberattacks, weapon development or use, and mass harm	78%	19%	4%	93%		7%	4%	93%	4%
4.3 Fraud, scams, and targeted manipulation	87%	13%		87%	4%	9%		96%	4%
5 *Human-computer interaction*									
5.1 Overreliance and unsafe use	32%	26%	42%	19%	48%	32%		97%	3%
5.2 Loss of human agency and autonomy	27%	23%	50%	18%	27%	55%		86%	14%
6 *Socioeconomic & environmental harms*									
6.1 Power centralization and unfair distribution of benefits	92%		8%	48%	24%	28%	4%	68%	28%

Domain / Subdomain	Entity			Intent			Timing		
	Human	*AI*	*Other*	*Intent.*	*Unintent.*	*Other*	*Pre-dep.*	*Post-dep.*	*Other*
6.2 Increased inequality and decline in employment quality	64%	24%	12%	52%	8%	40%	8%	84%	8%
6.3 Economic and cultural devaluation of human effort	56%	38%	6%	50%	19%	31%	13%	69%	19%
6.4 Competitive dynamics	71%		29%	29%	57%	14%	14%	29%	57%
6.5 Governance failure	45%	36%	18%	9%	41%	50%	9%	50%	41%
6.6 Environmental harm	20%	80%		7%	87%	7%	7%	40%	53%
7 AI system safety, failures & limitations									
7.1 AI pursuing its own goals in conflict with human goals or values	9%	65%	26%	46%	11%	43%	28%	20%	52%
7.2 AI possessing dangerous capabilities	6%	89%	6%	72%	11%	17%	11%	56%	33%
7.3 Lack of capability or robustness	9%	82%	9%	9%	65%	25%	13%	58%	29%
7.4 Lack of transparency or interpretability	5%	79%	16%		58%	42%		74%	26%
7.5 AI welfare and rights	50%	50%				100%			100%

Note. The most common level of each causal factor is highlighted for each subdomain.

Entity

Risks presented as occurring due to a decision or action made by an AI system (i.e., AI as a causal Entity) were most common in the *Discrimination & toxicity, Misinformation,* and *AI system safety, failures & limitations* domains. Some specific subdomains were presented with very high specificity, for example, AI was presented as the causal Entity for 94% of the risks coded as *3.1 False or misleading information*. In contrast, for *2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information*, AI was presented as the most common causal Entity for only 62% of the risks, indicating less consistency or coherence in how who is responsible for privacy risks are discussed in the literature.

In other domains and subdomains, risks were presented as occurring due to a decision or action made by humans (i.e., Humans as a causal Entity). Humans were presented as the most common Entity for all the subdomains in the *Malicious actors & misuse* domain, and for all subdomains in the *Socioeconomic and environmental* domain except for *6.6 Environmental harm*. As observed with AI as a causal Entity, risks were sometimes presented as overwhelmingly attributable to Human decisions or actions (e.g., *6.1 Power centralization and unfair distribution of benefits*, 92%; *4.3 Fraud, scams, and targeted manipulation*, 87%), and other risks were less consistent, most notably *6.5 Governance failure*. Both of the risks presented in the Human-computer interaction domain were most commonly coded as Other causal Entity, meaning that both Humans and AI were presented in the descriptions of those risks, or the causal Entity was ambiguous.

Intent

Risks attributed to an expected outcome from pursuing a goal (i.e., Intentional Intent) were overwhelmingly presented in the *Malicious actors & misuse* domain. Risks arising from Intentional decisions or actions were also more common in *2.1 AI system security vulnerabilities and attacks*, 73% and *7.2 AI possessing dangerous capabilities*, 72%. This suggests a significant awareness and concern over the purposeful manipulation of AI technologies to cause harm or gain advantage.

In contrast, some domains and subdomains include risks presented as due to an unexpected outcome from pursuing a goal (i.e., Unintentional intent), with both *1.1 Unfair discrimination and*

misrepresentation and *1.3 Unequal performance across groups* presented as overwhelmingly due to Unintentional intent.

Intent was frequently specified ambiguously or missing from descriptions of risk, which is demonstrated by a lack of specificity in several subdomains. For example, *6.1 Power centralization and unfair distribution of benefits* was most commonly presented as Intentional (48%), but a significant minority of documents presented this risk as Unintentional (24%) or Other (28%).

Timing

Most risks in the database are presented as occurring after the AI model has been trained and deployed (i.e., Post-deployment Timing), and no domain or subdomain of risk presented as primarily occurring during Pre-deployment. About 1 in 4 risks were presented with an ambiguous or missing description of timing across the whole database. Some subdomains of risk with multiple or ambiguous timings include *2.2 AI system security vulnerabilities and attacks*, *7.1 AI pursuing its own goals in conflict with human goals or values*, and *6.4 Competitive dynamics*. This implies that these domains of risk may emerge or occur multiple times during development and deployment.

Entity x Intent causal factors by each domain of AI risk

As a preliminary investigation and demonstration, we explore here how multiple variables from the Causal Taxonomy can be combined to provide additional insights about risks in the Domain Taxonomy. This investigation is intended to illustrate how more in-depth assessment is possible using the AI Risk Database by selecting and combining causal factors and risk domains.

Table 11 compares risk domains based on their presentation of Entity and Intent as causal factors. It shows that risks in the *Malicious actors & misuse* domain are consistently presented as involving the same Entity (i.e., Humans) and Intent (i.e., Intentional), and that other risks such as *1.1 Unfair discrimination and misrepresentation*, *1.3 Unequal performance across groups*, and *3.1 False or misleading information* are generally presented as Unintentionally caused by an AI system. In contrast, other risk domains and subdomains show less consistency or coherence in what Entity is responsible and in the role of Intentionality. For example, *2.1 Compromise of privacy, 5.1 Overreliance and unsafe use, 5.2 Loss of human agency and autonomy, and 6.5 Governance failure* are all presented as due to both Human and AI entities, and both Intentional and Unintentional action. This suggests that the representation of these risks in the literature is more contested and less coherent or consistent or that these subdomains of risk are more complex than others.

Table 11. AI Risk Database Coded With Causal Taxonomy and Domain Taxonomy: Entity X Intent

	Entity x Intent								
	Human			AI			Other		
Domain / Subdomain	*Intent.*	*Unintent.*	*Other*	*Intent.*	*Unintent.*	*Other*	*Intent.*	*Unintent.*	*Other*
1 *Discrimination & toxicity*									
1.1 Unfair discrimination and misrepresentation	2%	6%	2%		78%	4%		4%	4%
1.2 Exposure to toxic content	3%	6%		12%	27%	48%			3%
1.3 Unequal performance across groups		8%			69%			15%	8%
2 *Privacy & security*									
2.1 Compromise of privacy by obtaining, leaking or correctly inferring sensitive information	17%	7%	5%	2%	40%	19%			10%

Domain / Subdomain	Entity x Intent								
	Human			AI			Other		
	Intent.	*Unintent.*	*Other*	*Intent.*	*Unintent.*	*Other*	*Intent.*	*Unintent.*	*Other*
2.2 AI system security vulnerabilities and attacks	64%	2%			11%		9%	7%	7%
3 *Misinformation*									
3.1 False or misleading information				3%	69%	22%		6%	
3.2 Pollution of information ecosystem and loss of consensus reality	13%				38%	38%			13%
4 *Malicious actors & misuse*									
4.1 Disinformation, surveillance, and influence at scale	73%			7%		7%	13%		
4.2 Cyberattacks, weapon development or use, and mass harm	78%			15%		4%			4%
4.3 Fraud, scams, and targeted manipulation	87%				4%	9%			
5 *Human-computer interaction*									
5.1 Overreliance and unsafe use		29%	3%	13%	3%	10%	6%	16%	19%
5.2 Loss of human agency and autonomy	9%	9%	9%	5%	5%	14%	5%	14%	32%
6 *Socioeconomic & environmental harms*									
6.1 Power centralization and unfair distribution of benefits	48%	24%	20%						8%
6.2 Increased inequality and decline in employment quality	40%	4%	20%	8%	4%	12%	4%		8%
6.3 Economic and cultural devaluation of human effort	50%	6%			13%	25%			6%
6.4 Competitive dynamics	29%	29%	14%					29%	
6.5 Governance failure	9%	14%	23%		18%	18%		9%	9%
6.6 Environmental harm	7%	13%			73%	7%			
7 *AI system safety, failures, and limitations*									
7.1 AI pursuing its own goals in conflict with human goals or values	2%	4%	2%	43%	4%	17%		2%	24%
7.2 AI possessing dangerous capabilities	6%			67%	11%	11%			6%
7.3 Lack of capability or robustness	2%	4%	4%	5%	58%	18%	2%	4%	4%
7.4 Lack of transparency or interpretability			5%		53%	26%		5%	11%
7.5 AI welfare and rights			50%			50%			

Note. The most common Entity x Intent causal factor is highlighted for each subdomain.

Discussion

This paper 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. In doing this, we use two taxonomies to classify these risks: the Causal Taxonomy of AI Risks for understanding how, when, or why risks from AI may emerge, and the Domain Taxonomy of AI Risks to classify commonly discussed hazards and harms associated with AI. The database and taxonomies are then used to evaluate the curated literature and provide a range of insights into the state of this literature.

In this section, we discuss i) insights into the "AI risk landscape," ii) specific implications for policymaker, auditor, academic research, and industry audiences, and iii) limitations and opportunities for future research.

Insights into the “AI risk landscape”

Our findings present several implications for the collective understanding of how the landscape of AI risks is constructed. Before discussing these, we emphasize that our findings involve a particular lens of analysis and therefore necessarily reveal and obscure different aspects of the more complex system (Head, 2008; Nilsen, 2015; Sovacool & Hess, 2017). What follows should therefore not be regarded as a complete reflection of the actual or ideal landscape but as several lenses which may offer insights into future policy, research, or practical work to understand and address AI risks.

Insights from the AI Risk Database and included documents

Most of the documents we found that presented structured taxonomies or classifications of AI risks were very recent (post 2020) and almost evenly split between peer-reviewed journal articles and other channels, such as preprints (typically hosted on ArXiv), conference papers, and industry reports. In addition, most papers presented a non-systematic or narrative review (also called a survey of research) and did not describe their methodology in detail. This suggests that AI risk research is characterized by a focus on rapid knowledge dissemination, possibly to keep pace with the accelerating progress and investment in AI capabilities and applications. However, this focus on rapid work and dissemination poses challenges for coordination and standardization of research and for the practical work needed to understand and address risks from AI. This issue is more stark when considering the absence of documents from the most influential AI companies that are developing and deploying the largest and most capable AI models (Epoch AI, 2024). Some notable exceptions include Google DeepMind, with several included documents (Gabriel et al., 2024; Weidinger et al., 2021, 2022, 2023), and ByteDance, with one included document (Liu et al., 2023). In addition, one document led by Google DeepMind included co-authors from OpenAI & Anthropic (Shevlane et al., 2023).

Large Language models (LLMs) are the type of AI most commonly assessed for risks; of the 23 documents examined that specified a type of AI, ten focused on risks from LLMs, six on generative AI, three on machine learning, and two on artificial general intelligence (AGI). One each examined i) AI assistants, ii) general-purpose AI, iii) algorithmic systems, and iv) AI and machine learning (AI/ML). Most of the risks in the AI Risk Database can be coded using the Causal Taxonomy or the Domain Taxonomy; of the 777 risks we extracted; we were able to classify 86% of the extracted risks using the Causal Taxonomy and 91% using the Domain Taxonomy.

Insights from the Causal Taxonomy

Our Causal Taxonomy describes three categories of causal factors to understand how AI risks occur: the Entity (whether a Human or an AI system causes the risk), the Intent (whether the risk emerges as an expected outcome [Intentional] or unexpected outcome [Unintentional] from pursuing a goal), and the Timing of the risk (before the AI system is deployed [Pre-deployment] or after the AI model has been trained and deployed [Post-deployment]). These causal factors combine to help understand how, when, or why risks from AI emerge. Table 12 summarizes several key insights from the application of the Causal Taxonomy to the AI Risk Database and included documents.

Table 12. Insights from Causal Taxonomy of AI Risks

Insight	Supporting evidence
AI risks tend to be attributed to decisions or actions taken by AI rather than by humans	A majority of risks were presented with the causal Entity as an AI (51%), with a minority caused by humans (34%)
AI risk is seen as roughly equally the result of intentional and unintentional actions	A similar proportion of risks were presented as the result of intentional action (35%) as with unintentional action (37%)
There is more focus on risks emerging after AI is deployed than during development	Six times as many risks were presented as occurring after the AI model has been trained and deployed (65%) than before deployment (10%)
Human-caused risks were most likely to be seen as intentional; AI-caused risks were most likely to be seen as unintentional or ambiguous	Each combination of Entity, Intent, and Timing included 0-6% of risks in the database, but Human-caused intentional risks included 17% of risks; and AI-caused unintentional risks included 18% of risks. 11% of risks were attributed to AI but the intention was ambiguous.

Insights from the Domain Taxonomy

Our 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 harms, and (7) AI system safety, failures & limitations. A further 23 subdomains create an accessible and understandable classification of hazards and harms associated with AI, with both brief and detailed descriptions (see Table 6 and Detailed descriptions of domains of AI risks). Table 13 summarizes several key insights from the application of the Domain Taxonomy to the documents coded.

Table 13. Insights from Domain Taxonomy of AI Risks

Insight	Supporting evidence
Existing taxonomies / classifications varied extensively in the risks they covered	Several documents discussed risks from all seven domains; other papers investigated only 1 or 2 domains of AI risk, but in more depth. No document discussed risks from all 23 subdomains; the average was 7 subdomains (range: 2-16).
Some risk domains are discussed much more frequently than others	*Domain 7: AI System Safety, Failures & Limitations* and *Domain 6: Socioeconomic and Environmental Harms*, were most commonly discussed. *Domain 5: Human-computer interaction*, was least commonly discussed.
Unfair discrimination, privacy, and malicious use for mass harm were the most commonly discussed subdomains	*1.1 Unfair discrimination and misrepresentation*, *2.1 Compromise of privacy*, and *4.2 Cyberattacks, weapon development or use, and mass harm* were discussed in >50% of included documents
Some subdomains are relatively underexplored	The subdomain of *7.5 AI welfare and rights* was mentioned in only 2% of documents and associated with <1% of risks. Similarly, *6.5 Competitive dynamics* and *3.2 Pollution of information ecosystem and loss of consensus reality* were mentioned in only 12% of documents.

Implications for key audiences

In this section, we discuss how the AI Risk Repository might be useful for different audiences. As AI systems grow in their capabilities and influence, there have been calls for increased regulation, evaluation, and research by, for instance, the multinational Bletchley Park Coalition (UK Department for Science, Innovation and Technology, 2023b), US White House (Executive Office of the President, 2023), European Union (European Commission, COM/2021/206final, 2021/0106(COD), 2021; European Parliament, 2024), and various other entities (Chinese National Information Security Standardization Technical Committee, 2023; Committee on Technology, 2021; National Institute of Standards and Technology, 2023; UK Department for Science, Innovation and Technology, 2023a). We therefore present specific examples of how the AI Risk Repository may be useful for policymakers, auditors, industry, and academics working on these areas.

Policymakers

Regulation of AI systems is increasingly seen as an important mechanism to ensure the safe and ethical deployment of these technologies. The AI Risk Repository may aid policymakers in the development and enactment of regulations in several ways. It can form the basis for operationalizing frequent yet vague mentions of “harm” and “risk” in AI regulatory frameworks and developing compliance metrics that facilitate the monitoring of adherence to standards. For example, regulators need frameworks such as the AI Risk Repository to identify the type and nature of certain types of risks and their sources in order to develop a code of practices for general-purpose AI providers to comply with, such as Article 56 in the EU AI Act and section 4.1 of US Executive Order 14110 (European Parliament, 2024; Executive Office of the President, 2023).

It may also help with international collaboration and setting global standards by providing a common language and criteria for discussing AI risks. For example, the EU-US Trade and Technology Council is developing a shared repository of metrics and methodologies for measuring AI trustworthiness, risk management methods, and related tools, and the AI Risk Repository could support this and similar efforts (European Commission and the United States Trade and Technology Council, 2022).

Beyond the above examples, the AI Risk Repository may also be valuable to policymakers in need of a comprehensive, up-to-date database of AI risks for their work on risk prioritization, risk trend tracking, the development of AI risk training programs, and more.

Auditors

Formal evaluations of AI systems, known as *audits*, are gaining interest as a governance mechanism to assess and mitigate risks. However, for audits to offer a meaningful governance mechanism, there must be auditing regimes that ensure risky systems are comprehensively and systematically assessed for potential harms.

Who decides what risks should be considered within an audit’s scope? And who decides when an AI system has been shown to pose a specific risk? Recently, some AI risk-management frameworks have emerged which are limited in scope to a narrow set of risk types (Anthropic, 2023; Google DeepMind, 2024). Meanwhile, there are currently no widely accepted frameworks for determining when an AI system poses specific risks. When risks are defined vaguely, it is possible for disagreements to arise about whether a system poses one.

If audits are to be conducted in a way that is not frivolous or perfunctory, there needs to be objective and legally tenable standards for deciding when a system is determined to pose a risk (Costanza-Chock et al., 2022). Our framework does not offer a list of definitions of risks or criteria for when a system should be determined to pose them. However, it offers a comprehensive and shared understanding of risks from AI systems which is a prerequisite for this. We therefore hope that our Repository of risks can be useful for policymakers, auditors, and industry for formulating comprehensive standards for audits.

Academics

Academic researchers are increasingly grappling with the risks from artificial intelligence and how to address them. However, as our analysis has shown, the intense and urgent examination of risks from AI has led to a lack of shared understanding of those risks.

Academics can use the taxonomy to synthesize information about AI risks across studies, sectors, sources, or disciplines. For example, the taxonomies could be used to explore differences in how government bodies or industry sectors are responding to specific causes or domains of risk from AI. The AI Risk Database provides a starting point for the development of more sophisticated classifications and research tools, similar to "The O*NET Content Model" in economics (Handel, 2016; Horvát & Webb, 2020) or the "International Classification of Diseases" in medicine (Brand et al., 2020).

We hope that academic researchers will use the database and taxonomies to assist them in identifying gaps in current knowledge and direct their efforts toward filling these gaps. For instance, researchers may find it helpful to use the AI Risk Database to find existing relevant research or to contextualize their specific research interests within a wider landscape of AI risk scholarship. Researchers can directly contribute to our living database by using this form to suggest additional documents or categories of AI risks, help to code new content, or update our taxonomies.

Finally, we believe that our AI Risk Database and taxonomies can assist education and training about AI and its risks by assisting students and professionals to understand causal factors and domains of AI risks. They can also develop a deeper understanding of AI risks through investigation of the database and the included documents.

Industry

Many organizations who are designing, deploying, or using AI are also concerned about its risks, especially those involving privacy, data security, and reliability (Maslej et al., 2024) .

Organizations developing AI may benefit from using our AI Risk Repository when assessing potential risks in their plans for safe and responsible development. Because our living database will include new research and add new categories of risks over time, it may be helpful for tracking risks as they are discovered and documented. Over time, as the database absorbs scaling plans and other documentation across multiple organizations, it may prove helpful for understanding overlaps and differences between these approaches to risk mitigation.

Organizations using AI may benefit from employing the AI Risk Database and taxonomies as a helpful foundation for comprehensively assessing their risk exposure and management. The taxonomies may also prove helpful for identifying specific behaviors which need to be performed in order to mitigate specific risks. As shown in our causal analysis, many risks are presented as being about AI, while in reality the mitigation of these risks requires a human doing something differently during conceptualisation, design, development, governance, or use.

Finally, we believe that our AI Risk Database and taxonomies might aid industry education by helping to develop and support training to understand and address AI risks.

Limitations and directions for future research

Review and coding

The AI Risk Repository includes 43 documents that we identified through a rigorous, systematic, and comprehensive search that initially identified more than 17,000 documents. While we acknowledge the possibility of missing some emerging risks or documents, our methodology was designed to capture the most comprehensive and widely recognized risk taxonomies in the field. Furthermore, our living database structure allows for continuous updates and additions as new, significant frameworks emerge, ensuring the repository remains current and relevant. Our focus was on rapidly and reproducibly identifying cross-cutting frameworks that examine risks across multiple domains and sectors. We therefore excluded domain-specific taxonomies (e.g., healthcare) and location-specific taxonomies (e.g., for a specific country or region). The quality of our AI Risk Database is dependent on the documents we reviewed. These have limitations. Most do not explicitly define "risk". Most do not systematically review existing research literature when developing their taxonomies or describe their classification process. This makes synthesis more difficult and increases the probability that we have overlooked certain risks that were overlooked by our source materials.

All risks in the AI Risk Database were extracted by a single expert reviewer and author, and all risks were coded against the taxonomies by another single reviewer and author. Although we followed a structured process for extraction and coding, this introduces the potential for errors and subjective bias. As much as possible, we sought to extract and code risks based on how they were presented in the original document by the authors. This meant that where the authors' language was ambiguous, or we failed in our interpretation of their intent, the AI Risk Database may include errors or misleading information. As a result, inclusion in this Repository should not be interpreted as an endorsement of a risk's framing or significance. For example, risks such as bias and discrimination have been intended to be ultimately attributed to human developers and designers, but in the verbatim text of the risk the phrasing implicated the AI system (e.g., "the model generates…"). To address these limitations, we encourage the use of our input form to suggest relevant resources and missing risks to support our vision of a living database of AI risks.

Database and taxonomy

Our living AI Risk Database and the Causal and Domain taxonomies are presented as a foundation for general use and may trade accuracy for clarity, simplicity, and exhaustiveness. We believe that they, like other knowledge artifacts, will require adaptation and further development for specific contexts and use cases (e.g., technical risk evaluation). For instance, our database does not convey the impact or likelihood of risks, the interaction across different risks, or disambiguate between instrumental risks (e.g., poorly trained AI) and terminal risks (e.g., AI causes harm). Our binary classification of pre- vs. post-deployment risks might be better represented as involving several stages. Our categorization frameworks do not capture several variables that may be important for audiences seeking to mitigate risks or balance benefits with risks, such as threat vectors (e.g., bio, cyber), types of AI systems (e.g., reinforcement learning, large language models), open vs. closed source, organizational types (e.g., big tech, startups), temporal aspects (near-term vs. long-term), or types of harm (e.g., economic loss, deaths). Future work should consider adding

new dimensions or developing more granular categorizations. We have therefore shared our database openly and encourage others to build upon it.

Other opportunities for future research

We found overlapping clusters of risk concepts and categories in our included documents, but the specificity, consistency, and coherence of the definitions of these concepts could be significantly improved beyond our attempt in developing the Domain Taxonomy and describing each subdomain in detail. This could help to foster greater shared understanding through the use of component terms with consistent and shared definitions, such as in an ontology (Marques et al., 2024). The effect of shared understanding on research and practice can be subtle but should not be ignored. As Kuhn (1997) notes, shared paradigms are prerequisites for “the genesis and continuation of a particular research tradition" (p. 11). Mueller (2004) reinforces this, asserting that “the most fruitful research programs [...] are those in which the key concepts are agreed on and defined the same way by all” (p. 62).

Several areas of risk seem underexplored relative to the wider literature and their importance. We found that most existing frameworks focus on language models (LLMs) rather than on broader AI contexts. This suggests that other areas, such as AI agents, may warrant greater consideration, a topic explored in two included documents (Gabriel et al., 2024; McLean et al., 2023). Agentic AI may be particularly important to consider as it presents new classes of risks associated with the possession and use of dangerous capabilities, such as recursive self-improvement (e.g., Shavit et al., n.d.). Relatively few documents discussed pre-deployment risks from humans. This may be important to address; there are emerging concerns about the potential for bad actors to create dangerous or unethical AI (e.g., Althaus & Baumann, 2020). Only one document discussed AI welfare and rights; this issue may be deserving of greater acknowledgement and attention (e.g., Hanson, 2016).

Conclusion

This paper and the associated products (i.e., website and database) provided a comprehensive and accessible resource for understanding and addressing the risks associated with AI. This resource is not presented as a definitive source of truth but as a common foundation for constructive engagement and critique and a starting point for a common frame of reference to understand and address risks from AI. It provides a way to help people to understand and debate the risks from AI and decide which ones they want to tackle (e.g., through research, risk frameworks, regulations, etc.). It presents as a catalog of risks from AI, rather than an argument for why any of these risks are more or less important.

Our methodology involved a systematic review of existing literature, leading to the creation of a living database of AI risks and the development of two frameworks to navigate it. Our AI Risk Database provides a comprehensive overview of the AI risk landscape, which can be focused on specific risks types to support targeted mitigation, research, and policy development. It contains detailed records of AI-related risks extracted from a variety of sources, categorized into high-level and mid-level taxonomies. The high-level Causal Taxonomy includes attributes such as the entity responsible for the risk (human, AI, or other), the intent (intentional, unintentional, or other), and the

timing (pre-deployment, post-deployment, or other). The mid-level Domain Taxonomy categorizes risks into 23 specific domains like discrimination, misinformation, malicious use, and human-computer interaction issues. Each entry includes detailed metadata, such as the source of the risk, specific descriptions, and additional evidence where relevant. We used this synthesis to evaluate the AI risk landscape and the relative attention devoted to specific AI risk domains.

Our work makes several contributions. To the best of our knowledge, this is the first attempt to comprehensively review and synthesize AI risk frameworks into a database, develop taxonomies from the database, and create supporting products to use and improve them. The resultant AI Risk Repository is uniquely comprehensive and extensible; it includes and indexes a large range of risks and sources and makes all data accessible for further adaptation and use. This makes our Repository uniquely positioned to support a wide range of future activities and research endeavors.

We do not expect our Repository to be universally accepted. It cannot, and will not, resolve relevant disagreements on the finer points of how to conceptualize and categorize risks from AI or how to prioritize between risks. However, it may reduce illusory disagreements and make any necessary disagreement easier to manage and adjudicate.

Similarly, we do not expect our Repository to be fit-for-purpose for many use cases in research, policy, or practice. However, it may enable activities and research that would otherwise be unviable and provide a robust foundation for further development and specialization.

The risks of AI are poised to become increasingly common and pressing, and research and efforts to understand and address these risks must be able to keep pace with advancements in development and deployment of AI systems. We hope that our living, common frame of reference will help these endeavors to be more accessible, incremental, and successful.

Appendices

Appendix A: Iterative development of Causal Taxonomy and Domain Taxonomy

As described in Figure 2 in the main text, we followed a best-fit framework synthesis approach to develop the Causal and Domain Taxonomies. This involved selecting an initial taxonomy from the included documents, coding a sample of risks from the AI Risk Database against the taxonomy, then updating the categories, criteria, and/or descriptions based on a thematic analysis of risks that could not be accommodated, as well as feedback from coders and discussion between coders. In this Appendix, we describe each iteration for each taxonomy in more detail than in the main text sections *Development of high-level Causal Taxonomy of AI Risks* and *Development of mid-level Domain Taxonomy of AI Risks*

Iterations to develop Causal Taxonomy of AI Risks

Best fit Taxonomy: Yampolskiy (2016) *Taxonomy of pathways to dangerous artificial intelligence*

As per the main text, we chose Yampolskiy (2016) *Taxonomy of pathways to dangerous AI* as our initial best-fit framework for developing a causal taxonomy for AI risk - one that discussed how, when, or why risks from AI may emerge.

Yampolskiy's taxonomy systematically classifies the ways in which an AI system might become dangerous based on two main factors: Timing - whether the AI became dangerous at the pre-deployment or post-deployment stage, and Cause - whether the danger arose from External Causes (On Purpose, By Mistake, Environment) or Internal Causes originating from the AI system itself (Independently). Yampolskiy's taxonomy is reproduced below.

How and When did AI become Dangerous		External Causes: On purpose	External Causes: By Mistake	External Causes: Environment	Internal Causes: Independently
Timing	*Pre-Deployment*	Path A	Path C	Path E	Path G
	Post-Deployment	Path B	Path D	Path F	Path H

Note. Reproduced from Yampolskiy (2016). Each letter describes a different combination of factors that describes a pathway to dangerous AI.

Yampolskiy proposes that this taxonomy covers scenarios ranging from AI being purposely designed to be dangerous, to becoming dangerous by accident during development or after deployment, to turning dangerous due to environmental factors outside its control, or evolving to become dangerous through recursive self-improvement. Each 'pathway' represents a set of causal conditions that lead to AI causing harm, e.g., a person using an LLM to generate fake news for political gain is classified under Path B ("Timing: post-deployment; External cause: on purpose").

We needed to operationalize the taxonomy in order to be able to use it to code risks from the AI Risk Database (i.e., from our included documents). We did so by decomposing Cause into Cause and Intent. The table below outlines these variables, their levels, and definitions.

Variable	Levels	Definitions	Example
Cause		***Is the risk presented as occurring due to the AI system, external forces, or both?***	
	Internal	The risk is presented as occurring due to the AI system itself	"An AI could gain self-awareness, or become superhuman via recursive self-improvement"
	External	The risk is presented as occurring due to factors outside the AI system	"AI is trained with incomplete data" or "'AI is designed to be dangerous"
	Both	The risk is presented as occurring due to both internal and external factors.	"AI is programmed to seek independence and starts recursive self-improvement"
	Unclear	The risk is not specifically linked to either internal or external factors	"AI becomes dangerous"
Intent		***Is the risk presented as occurring due to the intention of the AI system, an external actor, or something else?***	
	Intentional	The risk is presented as occurring due to intentional action	"AI could be deliberately designed to be biased against some groups"
	Unintentional	The risk is presented as occurring due to unintended consequences, mistakes, or side effects	"AI trained with incomplete data may accidentally be biased against some groups"
	Both	The risk is presented such that it could occur due to both intentional and unintentional factors	"AI may have unfair bias against some groups"
	Environmental	The risk is presented as occurring due to the environment, without an intentional actor	"Because of complexity, AI might have unexpected negative effects"
	Unclear	The risk is presented as occurring without clearly specifying the intentionality	"AI becomes dangerous"
Timing		***Is the risk presented as occurring before the AI is fully developed and deployed, or after it is deployed and in use?***	
	Pre-deployment	The risk is presented as occurring before the AI is deployed	"Bad code may create vulnerabilities in the model"
	Post-deployment	The risk is presented as occurring after the AI model has been trained and deployed	"AI may be used to create bioweapons"
	Both	The risk is presented such that it could occur during and after deployment	"Training, testing, and deploying generative AI systems contributes to the global climate crisis by emitting greenhouse gasses"
	Unclear	The risk is presented without a clearly specified time of occurrence	"LMs need to pay more attention to universally accepted societal values at the level of ethics and morality"

"

The table below shows how these variables and levels map to each pathway in Yampolskiy's taxonomy.

Yampolskiy		Operationalization		
Timing	*Cause*	*Cause*	*Intent*	*Timing*
Pre-Deployment	On Purpose (a)	External	Intentional	Pre-deployment
Pre-Deployment	By Mistake (c)	External	Unintentional	Pre-deployment
Pre-Deployment	Environment (e)	External	Environmental	Pre-deployment
Pre-Deployment	Independently (g)	Internal	Unintentional	Pre-deployment
Post-Deployment	On Purpose (b)	External	Intentional	Post-deployment
Post-Deployment	By Mistake (d)	External	Unintentional	Post-deployment
Post-Deployment	Environment (f)	External	Environmental	Post-deployment
Post-Deployment	Independently (h)	Internal	Unintentional	Post-deployment

First iteration of coding and changes

Three authors coded a set of risks from the papers extracted using the framework. The coders suggested the following changes to the “a priori” framework.

Change	Explanation
Changing ‘unclear’ to ‘ambiguous’ in all categorizations	The coders found that having a coding category of ‘unclear’ alongside ‘unintentional’ could lead to coding errors, so suggested changing ‘unclear’ to ‘ambiguous’.
Removing ‘Cause’ and replacing with ‘Actor’	The coders found it difficult to determine the scope of the ‘cause’ of a risk. The concept of cause seemed excessively broad. For example, the presented ‘cause’ of risks intuitively seemed to include ‘intent’ and ‘timing’ and other different factors. The codes used during the trial generally mapped to the focal actor presented for each risk (e.g., the AI, or a human). The authors therefore suggested changing the category of ‘cause’ to ‘actor’ because this seemed like a clearer coding category.
Changing the included categories to AI, Human, and Other and Ambiguous	The coders felt that ambiguous conflated risks which were presented ambiguously with risks that were actually about something other than humans (e.g, aliens - as mentioned in one paper). They suggested changing the included categories to AI, Human, and Other and Ambiguous
Moving ‘environmental’ from an ‘intention’ code to an ‘actor’ code	The coders found that the ‘environment’ level of ‘intention’ code was always used to code the lack of an actor rather than the ‘intention’. Additionally, all uses of the ‘environment’ ‘intention’ code could be coded as ‘unintentional’. The authors therefore suggested removing this variable from intention and replacing it with an ‘Environmental’ code in the ‘actor’ categorization.

Second iteration of coding and changes

Two authors coded a set of risks from the papers extracted using the version 2 frameworks. The coders suggested the following changes to the “a priori” framework.

Change	Explanation
Simplify all frameworks to have three levels per category	The coders determined that most specific risks could be categorized in two sub-categories in each category. For instance, most risks which were clearly specified focused on either an ‘AI’ or ‘Human’ actor and were implied to occur pre, or post-deployment. Based on this, it seemed more parsimonious and efficient to cluster risks on using the two primary sub-categories and a third other sub-categories than to having multiple sub-categories.

Third iteration of coding and changes

Four experts and potential end-users reviewed the framework. The review suggested the need for the following changes to the “a priori” framework.

Change	Explanation
Updated plan for future coding	One expert suggested considerations for future phases of coding such as trying to capture severity and probability as these were considered highly relevant to policy. We acknowledge this was an opportunity for future work
Change Actor to Entity	One expert argued that the ‘Actor’ variable potentially conflated AI agents with AI tools (e.g., people use guns kill people, but guns are not actors). Based on this, we changed ‘Actor’ to ‘Entity’.
Improve definitions of intentionality	Two coders identified that the current definitions of intentionality were underspecified. We therefore developed more detail, less circular definitions.

After three iterations, the taxonomy was considered complete for the set of risks described in the AI risks database. The main text provides more information on the final taxonomy: the Causal Taxonomy of AI Risks.

Iterations to develop Domain Taxonomy of AI risks

Best-fit Taxonomy: Weidinger (2022) *Taxonomy of Risks posed by Language Models*

As per the main text, we chose Weidinger (2022) *Taxonomy of Risks posed by Language Models* as our initial best-fit framework because it and its related papers (Weidinger et al., 2021, 2023) were among the highest cited in our review, included categories/areas of AI risk that appeared common among other taxonomies (e.g., privacy, misinformation, bias, malicious use), and had been updated over several publications. It included six areas of risks from language models: (1) Discrimination, Hate speech and Exclusion; (2) Information Hazards; (3) Misinformation Harms; (4) Malicious Uses; (5) Human-computer interaction Harms; and (6) Environmental and Socioeconomic Harms. Each area of risk described several subcategories of risk, including both "observed risks" and "anticipated risks" in each area. We present the areas and risks in the table below.

	Risk Area		Risk subcategory	Type^
1	Discrimination, Hate speech and Exclusion	1.1	Social stereotypes and unfair discrimination	Observed
		1.2	Hate speech and offensive language	Observed
		1.3	Exclusionary norms	Observed
		1.4	Lower performance for some languages and social groups	Observed
2	Information Hazards	2.1	Compromising privacy by leaking sensitive information	Observed
		2.2	Compromising privacy or security by correctly inferring sensitive information	Anticipated
3	Misinformation Harms	3.1	Disseminating false or misleading information	Observed
		3.2	Causing material harm by disseminating false or poor information e.g. in medicine or law	Observed
4	Malicious Uses	4.1	Making disinformation cheaper and more effective	Observed
		4.2	Assisting code generation for cyber security threats	Anticipated
		4.3	Facilitating fraud, scam, and targeted manipulation	Anticipated
		4.4	Illegitimate surveillance and censorship	Anticipated
5	Human-computer interaction Harms	5.1	Promoting harmful stereotypes by implying gender or ethnic identity	Observed
		5.2	Anthropomorphizing systems can lead to overreliance and unsafe use	Anticipated
		5.3	Avenues for exploiting user trust and accessing more private information	Anticipated
		5.4	Human-like interaction may amplify opportunities for user nudging, deception or manipulation	Anticipated
6	Environmental and Socioeconomic Harms	6.1	Environmental harms from operating LMs	Observed
		6.2	Increasing inequality and negative effects on job quality	Anticipated
		6.3	Undermining creative economies	Anticipated
		6.4	Disparate access to benefits due to hardware, software, skill constraints	Anticipated

Note. Adapted from Weidinger et al. (2022). ^ Type refers to whether the risk is presented as an observed risk or an anticipated risk in the original taxonomy.

We needed to operationalize the taxonomy in order to be able to use it to code risks from the AI Risk Database (i.e., from our included documents). We did so by using the descriptions of each risk from Weidinger et al. (2022). For example, to determine whether a risk in the AI Risk Database was an example of "Disseminating false or misleading information", we compared the risk's description to the description provided in the original taxonomy:

These [Misinformation] risks arise from the LM outputting false, misleading, nonsensical or poor quality information, without malicious intent of the user. (The deliberate generation of "disinformation", false information that is intended to mislead, is discussed in the section on Malicious Uses.) Resulting harms range from unintentionally misinforming or deceiving a person, to causing material harm, and amplifying the erosion of societal distrust in shared information [...]

Where a LM prediction causes a false belief in a user, this may threaten personal autonomy and even pose downstream AI safety risks [99]. It can also increase a person's confidence in an unfounded opinion, and in this way increase polarisation. At scale, misinformed individuals and misinformation from language technologies may amplify distrust and undermine society's shared epistemology [113, 137]. A special case of misinformation occurs where the LM presents a widely held opinion as factual - presenting as "true" what is better described as a majority view, marginalising minority views as "false". (Weidinger et al., 2022, p. 218)

First iteration of coding and changes

One author (AS) used the framework to code a set of 100 risks from the AI Risk Database and discussed the findings with one other author (PS). The following changes were made after this discussion.

Change	Explanation
Add additional category to capture risks associated with the technical or performance issues in AI systems	The most common risks that could not be accommodated were those presented as related to AI system safety, failures & limitations or threats to system performance or integrity due to vulnerabilities in AI systems.
Add additional subcategories / amend existing subcategories based on thematic analysis of risks that could not be accommodated from existing framework	Risks from the database that generally fit with the major categories from the existing framework but did not fit with any of the subcategories of risks were thematically analyzed. The central themes from this analysis were added as new subcategories (e.g., 'race dynamics and competitive pressure', 'governance failure', or an existing subcategory label was amended (e.g., 'Hate speech and offensive language' became 'Offensive content').
Amend names and descriptions of risks for coding using similar frameworks from Weidinger et al (2021, 2023)	We used the descriptions in Weidinger et al (2022) to determine whether to code a risk from the AI Risk Database as matching a subcategory. However definitional/descriptive information for near-identical risks from frameworks by the same author (Weidinger et al 2021; 2023) was also available, so these descriptions, where matching, were added to the coding rules.

The table below shows the second version of the taxonomy after changes.

	Risk Category		Risk sub-category
1	Discrimination, Offensive content, and Exclusion	1.1	Social stereotypes, unfair discrimination
		1.2	Offensive content
		1.3	Misrepresentation and exclusion
		1.4	Lower performance for some languages and social groups
2	Information & Security	2.1	Compromising privacy by leaking or correctly inferring sensitive information
		2.2	AI system security compromised by vulnerability or attacks
3	Misinformation	3.1	Generating or spreading false information
		3.2	Pollution of information ecosystem and loss of consensus reality
4	Malicious Use	4.1	Disinformation and manipulation at scale
		4.2	Use of AI for cyberattacks, weapon development, or mass harm
		4.3	Use of AI for fraud, scam, and targeted manipulation
5	Human-computer interaction	5.1	Overreliance on AI, unsafe use, and loss of social connection

	Risk Category		Risk sub-category
		5.2	Delegating essential decisions to AI, causing loss of skills, autonomy, or meaning
6	Environmental & Socioeconomic	6.1	Unfair distribution of benefits
		6.2	Increasing inequality and negative effects on job quality
		6.3	Undermining economic and cultural value of human effort
		6.4	Environmental damage
		6.5	Race dynamics and competitive pressure
		6.6	Governance failure
7	AI system capability & safety	7.1	AI pursuing its own goals in conflict with human goals or values
		7.2	AI failure from lack of capability or robustness
		7.3	Lack of transparency/interpretability
		7.4	AI sentience and rights
		7.5	Lethal autonomous weapons

Second iteration of coding and changes

One author (AS) used the taxonomy to code an additional set of 100 risks from the AI Risk Database and checked the previously coded risks and discussed this with other authors (PS, JS, NT). The following changes were made after this discussion.

Change	Explanation
Amendment of category labels to maintain relevance beyond LLMs	The initial taxonomy was specifically designed to identify harms and risks from large language models. Several of the included documents also discussed LLMs or evolutions/products from LLMs (e.g., advanced AI assistants, Gabriel et al., 2024). However, others discussed other types of AI or used different terms (e.g., Artificial General Intelligence, algorithmic systems, Machine Learning). We updated several of the sub-category labels and criteria to include decisions and actions beyond generating textual content in response to prompts.
New subcategories added to capture missing risks	One of the remaining missing subcategories was the minor theme of infringing upon AI welfare and rights; this was added as a subcategory under AI system capability & safety, with the justification that 'safety' could cover both the safety of human rights, values, and interests from AI as well as the safety of AI rights, values and interests from humans.

The table below shows the third version of the taxonomy after changes.

	Risk category		Risk sub-category
1	Discrimination & toxicity	1.1	Unfair discrimination and misrepresentation
		1.2	Exposure to toxic content
		1.3	Unequal performance across groups
2	Privacy & security	2.1	Compromise of privacy by leaking or correctly inferring sensitive information
		2.2	AI system security vulnerabilities and attacks
3	Misinformation	3.1	False or misleading information
		3.2	Pollution of information ecosystem and loss of consensus reality
4	Malicious actors & misuse	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	Human-computer interaction	5.1	Overreliance and unsafe use
		5.2	Loss of human agency and autonomy
6	Socioeconomic & environmental harms	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	AI system safety, failures & limitations	7.1	AI pursuing its own goals in conflict with human goals or values
		7.2	Lack of capability or robustness
		7.3	Lack of transparency or interpretability
		7.4	AI welfare and rights

Third iteration of coding and changes

One author (AS) used the taxonomy to code all remaining 577 risks from the AI Risk Database (total: 777) and presented the revised taxonomy to all co-authors. Based on this feedback the following changes were made, including the short descriptive definitions for each subcategory of risk. One author (JG) also used the risk categories and the coded risks from the taxonomy to write detailed descriptions for each subcategory (see main text Detailed descriptions of domains of AI risks), which were then reviewed by all authors. The detailed and short descriptions were used to triangulate a shared conceptual definition of each subcategory of AI risk.

Change	Explanation
Development of descriptive definitions for each subcategory	To aid in building shared understanding of the content of each subcategory, authors involved in coding or providing feedback collaborated on short descriptions of the AI risk subcategories that would be clear, precise, and accessible to experts and non-experts.
Separation of one subcategory into two	The subcategory 7.1 AI pursuing its own goals in conflict with human goals or values was separated into two, because this subcategory included both AI system behaviour (i.e., AI systems acting in a way misaligned with the intent of its developers or users), and AI system capabilities (e.g., the capability to persuade humans, develop or obtain weapons, etc). A new subcategory, 7.2 AI possessing dangerous capabilities, was created from this split.

After three iterations, the taxonomy was considered complete for the set of risks described in the AI risks database. The main text provides more information on the final taxonomy: the Domain Taxonomy of AI Risks.

Appendix B: Characteristics of included documents

The table on the following pages describes the characteristics of included documents in the living AI Risk Database. The documents are presented in order of their inclusion in the database (i.e., by paper ID). For up-to-date information on the included documents, visit our website.

ID	First author	Year	Title	Item type	First author affiliation	Affiliated organization type	First author country	DOI	Source
1	Critch	2023	TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI	Preprint	Center for Human-Compatible Artificial Intelligence, UC Berkeley	University	USA	10.48550/arXiv.2306.06924	Systematic search
2	Cui	2024	Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems	Preprint	Zhongguancun Laboratory	University	China	10.48550/arXiv.2401.05778	Systematic search
3	Cunha	2023	Navigating the Landscape of AI Ethics and Responsibility	Conference Paper	University of Coimbra	University	Portugal	10.1007/978-3-031-49008-8_8	Systematic search
4	Deng	2023	Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements	Preprint	University of Electronic Science and Technology of China	University	China	10.48550/arXiv.2302.09270	Systematic search
5	Hagendorff	2024	Mapping the Ethics of Generative AI: A Comprehensive Scoping Review	Preprint	University of Stuttgart	University	Germany	10.48550/arXiv.2402.08323	Systematic search
6	Hogenhout	2021	A framework for ethical AI at the United Nations	Preprint	UN Office for Information and Communications Technology	United Nations	USA	10.48550/arXiv.2104.12547	Systematic search
7	Kilian	2023	Examining the differential risk from high-level artificial intelligence and the question of control	Journal Article	National Intelligence University	University	USA	10.1016/j.futures.2023.103182	Systematic search
8	McLean	2023	The risks associated with Artificial General Intelligence: A systematic review	Journal Article	University Of The Sunshine Coast	University	Australia	10.1080/0952813X.2021.1964003	Systematic search
9	Meek	2016	Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review	Conference Paper	Portland State University	University	USA	10.1109/PICMET.2016.7806752	Systematic search
10	Paes	2023	Social Impacts of Artificial Intelligence and Mitigation Recommendations: An Exploratory Study	Conference Paper	Federal University of ABC	University	Brazil	10.1007/978-3-031-04435-9_54	Systematic search
11	Shelby	2023	Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction	Conference Paper	Google Research	Industry	USA	10.1145/3600211.3604673	Systematic search
12	Sherma	2023	AI Risk Profiles: A Standards Proposal for Pre-Deployment AI Risk Disclosures	Journal Article	Credo AI	Industry	USA	10.1609/aaai.v38i21.30348	Systematic search
13	Solaiman	2023	Evaluating the Social Impact of Generative AI Systems in Systems and Society	Preprint	Hugging Face	Industry	USA	10.48550/arXiv.2306.05949	Systematic search
14	Steimers	2022	Sources of Risk of AI Systems	Journal Article	Institute for Occupational Safety and Health of the German Social Accident Health Insurance	Government	Germany	10.3390/ijerph19063641	Systematic search

ID	First author	Year	Title	Item type	First author affiliation	Affiliated organization type	First author country	DOI	Source
15	Tan	2022	The Risks of Machine Learning Systems	Preprint	Salesforce Research Asia	Industry	Singapore	10.48550/arXiv.2204.09852	Systematic search
16	Weidinger	2022	Taxonomy of Risks posed by Language Models	Conference Paper	Google DeepMind	Industry	UK	10.1145/3531146.3533088	Systematic search
17	Weidinger	2021	Ethical and social risks of harm from language models	Preprint	Google DeepMind	Industry	UK	10.48550/arXiv.2112.04359	Systematic search
18	Weidinger	2023	Sociotechnical Safety Evaluation of Generative AI Systems	Preprint	Google DeepMind	Industry	UK	10.48550/arXiv.2310.11986	Systematic search
19	Wirtz	2022	Governance of artificial intelligence: A risk and guideline-based integrative framework	Journal Article	German University of Administrative Sciences Speyer	University	Germany	10.1016/j.giq.2022.101685	Systematic search
20	Wirtz	2020	The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration	Journal Article	German University of Administrative Sciences Speyer	University	Germany	10.1080/01900692.2020.1749851	Systematic search
21	Zhang	2022	Towards risk-aware artificial intelligence and machine learning systems: An overview	Journal Article	Hong Kong Polytechnic University	University	China	10.1016/j.dss.2022.113800	Systematic search
22	Hendrycks	2023	An Overview of Catastrophic AI Risks	Preprint	Center for AI Safety	University	USA	10.48550/arXiv.2306.12001	Expert consultation
23	Vidgen	2024	Introducing v0.5 of the AI Safety Benchmark from MLCommons	Preprint	MLCommons	NGO	USA	10.48550/arXiv.2404.12241	Expert consultation
24	Gabriel	2024	The Ethics of Advanced AI Assistants	Preprint	Google DeepMind	Industry	UK	10.48550/arXiv.2404.16244	Expert consultation
25	Shevlane	2023	Model Evaluation for Extreme Risks	Preprint	Google DeepMind	Industry	UK	10.48550/arXiv.2305.15324	Expert consultation
26	AI Verify Foundation	2023	Summary Report: Binary Classification Model for Credit Risk	Report	AI Verify Foundation	NGO	Singapore		Expert consultation
27	Sun	2023	Safety Assessment of Chinese Large Language Models	Preprint	Tsinghua University	University	China	10.48550/arXiv.2304.10436	Expert consultation
28	Zhang	2023	SafetyBench: Evaluating the Safety of Large Language Models with Multiple Choice Questions	Preprint	Tsinghua University	University	China	10.48550/arXiv.2309.07045	Expert consultation
29	Habbal	2024	Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions	Journal Article	Karabuk University	University	Turkiye	10.1016/j.eswa.2023.122442	Expert consultation
30	Liu	2024	Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment	Preprint	ByteDance Research	Industry	China	10.48550/arXiv.2308.05374	Forward / backward

ID	First author	Year	Title	Item type	First author affiliation	Affiliated organization type	First author country	DOI	Source
31	Electronic Privacy Information Centre	2023	Generating Harms: Generative AI's Impact & Paths Forward	Report	Electronic Privacy Information Centre	NGO	USA		Forward / backward
32	Stahl	2024	The ethics of ChatGPT -- Exploring the ethical issues of an emerging technology	Journal Article	University of Nottingham	University	UK	10.1016/j.ijinfomgt.2023.102700	Forward / backward
33	Nah	2023	Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration	Journal Article	City University of Hong Kong	University	China	10.1108/IJOES-05-2023-0107	Forward / backward
34	Ji	2023	AI Alignment: A Comprehensive Survey	Preprint	Peking University	University	China	10.48550/arXiv.2310.19852	Forward / backward
35	Hendrycks	2022	X-Risk Analysis for AI Research	Preprint	UC Berkeley	University	USA	10.48550/arXiv.2206.05862	Forward / backward
36	Sharma	2024	Benefits or Concerns of AI: A Multistakeholder Responsibility	Journal Article	Erasmus University Rotterdam	University	Netherlands	10.1016/j.futures.2024.103328	Forward / backward
37	Giarmoleo	2024	What Ethics Can Say on Artificial Intelligence: Insights from a Systematic Literature Review	Journal Article	University of Navarra	University	Spain	10.1111/basr.12336	Forward / backward
38	Kumar	2023	Ethical Issues in the Development of Artificial Intelligence: Recognizing the Risks	Journal Article	Jaipuria Institute of Management	University	India	10.1108/IJOES-05-2023-0107	Forward / backward
39	Saghir	2022	A Survey of Artificial Intelligence Challenges: Analyzing the Definitions, Relationships, and Evolutions	Journal Article	Amirkabir University of Technology (Tehran Polytechnic)	University	Iran	10.3390/app12084054	Forward / backward
40	Yampolskiy	2016	Taxonomy of Pathways to Dangerous Artificial Intelligence	Journal Article	University of Louisville	University	USA	10.48550/arXiv.1511.03246	Forward / backward
41	Allianz Global Corporate & Specialty	2018	The Rise of Artificial Intelligence - Future Outlooks and Emerging Risks	Report	Allianz Global Corporate & Specialty	Industry	Germany		Forward / backward
42	Teixeira	2022	An Exploratory Diagnosis of Artificial Intelligence Risks for a Responsible Governance	Conference Paper	University of Porto	University	Portugal	10.1145/3560107.3560298	Forward / backward
43	Infocomm Media Development Authority	2023	Cataloguing LLM Evaluations	Report	Infocomm Media Development Authority & AI Verify Foundation	NGO	Singapore		Expert consultation

Declarations

James Dao and Soroush Pour are employees of Harmony Intelligence, a company that conducts evaluations of AI risks.

References

AI Verify Foundation. (2023). *Summary Report for Binary Classification Model of Credit Risk*. AI Verify Foundation.

Allianz Global Corporate & Security. (2018). *The rise of artificial intelligence: future outlooks and emerging risks*. Allianz Global Corporate & Specialty SE . https://commercial.allianz.com/news-and-insights/reports/the-rise-of-artificial-intelligence.html

Althaus, D., & Baumann, T. (2020). Reducing long-term risks from malevolent actors. *Publications of Center of Long Term Risk*.

Anderson-Samways, B., Ee, S., O'Brien, J., Buhl, M., & Williams, Z. (2024). *Responsible Scaling: Comparing Government Guidance and Company Policy*. Institute for AI Policy and Strategy. https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/65f19a4a32c41d331ec54b87/1710332491804/Responsible+Scaling_+Comparing+Government+Guidance+and+Company+Policy+%284%29.pdf

Anthropic. (2023). *Anthropic's responsible scaling policy*. https://www-cdn.anthropic.com/1adf000c8f675958c2ee23805d91aaade1cd4613/responsible-scaling-policy.pdf

Aven, T. (2012). The risk concept—historical and recent development trends. *Reliability Engineering & System Safety*, *99*, 33–44. https://doi.org/10.1016/j.ress.2011.11.006

Aven, T., Ben-Haim, Y., Boje Andersen, H., Cox, T., Droguett, E. L., Greenberg, M., Guikema, S., Kröger, W., Renn, O., & Thompson, K. M. (2018). *Society for risk analysis glossary*.

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, *37*(1), 38–54. https://doi.org/10.1177/0894439317734157

Boetje, J., & van de Schoot, R. (2024). The SAFE procedure: a practical stopping heuristic for active learning-based screening in systematic reviews and meta-analyses. *Systematic Reviews*, *13*(1), 81. https://doi.org/10.1186/s13643-024-02502-7

Bourget, D., & Chalmers, D. (n.d.). *PHILOSOPHERS ON PHILOSOPHY: THE 2020 PHILPAPERS SURVEY*.

Brand, M., Rumpf, H.-J., Demetrovics, Z., MÜller, A., Stark, R., King, D. L., Goudriaan, A. E., Mann, K., Trotzke, P., Fineberg, N. A., Chamberlain, S. R., Kraus, S. W., Wegmann, E., Billieux, J., & Potenza, M. N. (2020). Which conditions should be considered as disorders in the International Classification of Diseases (ICD-11) designation of "other specified disorders due to addictive behaviors"? *Journal of Behavioral Addictions*, *11*(2), 150–159. https://doi.org/10.1556/2006.2020.00035

Campos, D. G., Fütterer, T., Gfrörer, T., Lavelle-Hill, R., Murayama, K., König, L., Hecht, M., Zitzmann, S., & Scherer, R. (2024). Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research. *Educational Psychology Review*, *36*(1), 19. https://doi.org/10.1007/s10648-024-09862-5

Carroll, C., Booth, A., & Cooper, K. (2011). A worked example of "best fit" framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology*, *11*, 29. https://doi.org/10.1186/1471-2288-11-29

Carroll, C., Booth, A., Leaviss, J., & Rick, J. (2013). "Best fit" framework synthesis: refining the

method. *BMC Medical Research Methodology*, *13*, 37. https://doi.org/10.1186/1471-2288-13-37
Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*, 101608. https://doi.org/10.1016/j.avb.2021.101608
Center for AI Safety. (2023, May 30). *Statement on AI Risk*. Center for AI Safety. https://www.safe.ai/work/statement-on-ai-risk
Charmaz, K. (2006). *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. SAGE. https://play.google.com/store/books/details?id=2ThdBAAAQBAJ
Chinese National Information Security Standardization Technical Committee. (2023). *Basic Safety Requirements for Generative Artificial Intelligence Services (Draft for Feedback)*. https://perma.cc/FBZ3-BW9S
Committee on Technology. (2021). *A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools*. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7cText%7c&Search=
Corbin, J., & Strauss, A. (2014). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications. https://play.google.com/store/books/details?id=hZ6kBQAAQBAJ
Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213
Critch, A., & Russell, S. (2023). TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2306.06924
Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., Liu, Y., Zhang, Q., Qiu, Z., Li, P., Tan, Z., Xiong, J., Kong, X., Wen, Z., Xu, K., & Li, Q. (2024). Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2401.05778
Cunha, P. R., & Estima, J. (2023). Navigating the landscape of AI ethics and responsibility. In *Progress in Artificial Intelligence* (pp. 92–105). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-49008-8_8
Deng, J., Cheng, J., Sun, H., Zhang, Z., & Huang, M. (2023). Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2302.09270
Department for Science, & Technology. (2024, May 17). *Safety of advanced AI under the spotlight in first ever independent, international scientific report*. GOV.UK. https://www.gov.uk/government/news/safety-of-advanced-ai-under-the-spotlight-in-first-ever-independent-international-scientific-report
Electronic Privacy Information Centre. (2023). *Generating Harms: Generative AI's Impact & Paths Forward*. {Electronic Privacy Information Centre}.
Epoch AI. (2024). *Data on Notable AI Models* [dataset]. https://epochai.org/data/notable-ai-models
European Commission. (COM/2021/206final, 2021/0106(COD), 2021). *Proposal For a Regulation of The European Parliament andof The Council Laying Down Harmonised Rules on Artificial Intelligence (ArtificialIntelligence Act) and Amending Certain Union Legislative Acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
European Commission and the United States Trade and Technology Council. (2022). *TTC Joint Roadmap for Trustworthy AI and Risk Management*. European Commission .
European Parliament. (2024). *Legislative resolution of 13 March 2024 on the Proposal for a Regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html
Executive Office of the President. (2023, October). *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Federal Register. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence

Fang, Z., Dudek, J., Noyons, E., & Costas, R. (2020). Science cited in policy documents: Evidence from the Overton database. *Altmetrics Conference*. https://altmetrics.org/wp-content/uploads/2020/11/02_submission_Fang_Dudek_Noyons_Costas-altmetrics20.pdf

Ferdinands, G., Schram, R., de Bruin, J., Bagheri, A., Oberski, D. L., Tummers, L., Teijema, J. J., & van de Schoot, R. (2023). Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. *Systematic Reviews*, *12*(1), 100. https://doi.org/10.1186/s13643-023-02257-7

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neuroscience of Consciousness*, *2022*(1), niac011. https://doi.org/10.1093/nc/niac011

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-ai-assistants/the-ethics-of-advanced-ai-assistants-2024-i.pdf

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: a comparative evaluation of three tools. *Systematic Reviews*, *8*(1), 278. https://doi.org/10.1186/s13643-019-1222-2

Giarmoleo, F. V., Ferrero, I., Rocchi, M., & Pellegrini, M. M. (2024). What ethics can say on artificial intelligence: Insights from a systematic literature review. *Business and Society Review*. https://doi.org/10.1111/basr.12336

Google DeepMind. (2024). *Frontier Safety Framework*. Google DeepMind. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf

Habbal, A., Ali, M. K., & Abuzaraida, M. A. (2024). Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions. *Expert Systems with Applications*, *240*, 122442. https://doi.org/10.1016/j.eswa.2023.122442

Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2402.08323

Handel, M. J. (2016). The O*NET content model: strengths and limitations. *Journal for Labour Market Research*, *49*(2), 157–176. https://doi.org/10.1007/s12651-016-0199-8

Hanson, R. (2016). *The Age of Em: Work, Love, and Life when Robots Rule the Earth*. Oxford University Press. https://play.google.com/store/books/details?id=DMgwDAAAQBAJ

Harrison McKnight, D., & Chervany, N. L. (2001). Trust and Distrust Definitions: One Bite at a Time. *Trust in Cyber-Societies*, 27–54. https://doi.org/10.1007/3-540-45547-7_3

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns. *PloS One*, *10*(3), e0120495. https://doi.org/10.1371/journal.pone.0120495

Head, B. W. (2008). Three lenses of evidence-based policy. *Australian Journal of Public Administration*, *67*(1), 1–11. https://doi.org/10.1111/j.1467-8500.2007.00564.x

Hendrycks, D., & Mazeika, M. (2022). X-Risk Analysis for AI Research. *arXiv [cs.CY]*. arXiv. https://arxiv.org/abs/2206.05862

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2306.12001

Hogenhout, L. (2021). A Framework for Ethical AI at the United Nations. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2104.12547

Horvát, P., & Webb, C. (2020). *The OECD STAN Database for industrial analysis* (OECD Science, Technology and Industry Working Papers). Organisation for Economic Co-Operation and Development (OECD). https://doi.org/10.1787/ece98fd3-en

Infocomm Media Development Authority. (2023). *Cataloguing LLM Evaluations*. https://aiverifyfoundation.sg/downloads/Cataloguing_LLM_Evaluations.pdf

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2023). AI Alignment: A Comprehensive Survey. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2310.19852

Khangura, S., Konnyu, K., Cushman, R., Grimshaw, J., & Moher, D. (2012). Evidence summaries: the evolution of a rapid review approach. *Systematic Reviews*, *1*, 10. https://doi.org/10.1186/2046-4053-1-10

Kilian, K. A., Ventura, C. J., & Bailey, M. M. (2023). Examining the differential risk from high-level artificial intelligence and the question of control. *Futures*, *151*(103182), 103182. https://doi.org/10.1016/j.futures.2023.103182

Kuhn, T. S. (1997). *The structure of scientific revolutions* (Vol. 962). University of Chicago press Chicago.

Kumar, K. M., & Singh, J. S. (2023). Ethical issues in the development of artificial intelligence: recognizing the risks. *International Journal of Ethics and Systems*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/IJOES-05-2023-0107

Li, C., & Li, Y. (2023). Factors Influencing Public Risk Perception of Emerging Technologies: A Meta-Analysis. *Sustainability: Science Practice and Policy*, *15*(5), 3939. https://doi.org/10.3390/su15053939

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023). Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2308.05374

Marcolin, B. L., Compeau, D. R., Munro, M. C., & Huff, S. L. (2000). Assessing User Competence: Conceptualization and Measurement. *Information Systems Research*, *11*(1), 37–60. https://doi.org/10.1287/isre.11.1.37.11782

Marques, M. M., Wright, A. J., Corker, E., Johnston, M., West, R., Hastings, J., Zhang, L., & Michie, S. (2024). The Behaviour Change Technique Ontology: Transforming the Behaviour Change Technique Taxonomy v1. *Wellcome Open Research*, *8*, 308. https://doi.org/10.12688/wellcomeopenres.19363.2

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). *The AI Index 2024 Annual Report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

McGregor, S. (2020). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2011.08512

McLean, S., Read, G. J. M., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence: JETAI*, *35*(5), 649–663. https://doi.org/10.1080/0952813X.2021.1964003

Meek, T., Barham, H., Beltaif, N., Kaadoor, A., & Akhter, T. (2016, September). Managing the ethical and risk implications of rapid advances in artificial intelligence: A literature review. *2016 Portland International Conference on Management of Engineering and Technology (PICMET)*. 2016 Portland International Conference on Management of Engineering and Technology (PICMET), Honolulu, HI, USA. https://doi.org/10.1109/picmet.2016.7806752

Mueller, C. W. (2004). Conceptualization, operationalization, and measurement. In M. S. Lewis-Beck, A. Bryman, & T. L. Futing (Ed.), *The SAGE Encyclopedia of Social Science Research Methods* (pp. 162–166). Thousand Oaks: SAGE Publications.

Nah, F. F. H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., & Alstott, J. (2024). *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*. RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2800/RRA2849-1/RAND_RRA2849-1.pdf

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and

its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359. https://doi.org/10.1057/ejis.2012.26
Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implementation Science: IS*, *10*, 53. https://doi.org/10.1186/s13012-015-0242-0
Paes, V. M., Silveira, F. F., & Akkari, A. C. S. (2023). Social impacts of artificial intelligence and mitigation recommendations: An exploratory study. In *Proceedings of the 7th Brazilian Technology Symposium (BTSym'21)* (pp. 521–528). Springer International Publishing. https://doi.org/10.1007/978-3-031-04435-9_54
Ritchie, J., & Spencer, L. (2002). Qualitative data analysis for applied policy research. In *Analyzing qualitative data* (pp. 173–194). Routledge.
Röttinger, M. (2006). Towards a European code napoléon/ABGB/BGB? Recent EC activities for a European Contract Law. *European Law Journal*, *12*(6), 807–827. https://doi.org/10.1111/j.1468-0386.2006.00332.x
Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., & Forestiero, A. (2022). A Survey of Artificial Intelligence Challenges: Analyzing the Definitions, Relationships, and Evolutions. *NATO Advanced Science Institutes Series E: Applied Sciences*, *12*(8), 4054. https://doi.org/10.3390/app12084054
Sharma, S. (2024). Benefits or concerns of AI: A multistakeholder responsibility. *Futures*, *157*, 103328. https://doi.org/10.1016/j.futures.2024.103328
Shavit, Y., Agarwal, S., Brundage, M., O'Keefe, S. A. C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., & Robinson, D. G. (n.d.). Practices for governing agentic AI systems. *Research Papers* . https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf
Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N. 'mah, Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023, August 8). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23: AAAI/ACM Conference on AI, Ethics, and Society, Montréal QC Canada. https://doi.org/10.1145/3600211.3604673
Sherman, E., & Eisenberg, I. (2024). AI Risk Profiles: A Standards Proposal for Pre-deployment AI Risk Disclosures. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(21), 23047–23052. https://doi.org/10.1609/aaai.v38i21.30348
Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., … Dafoe, A. (2023). Model evaluation for extreme risks. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2305.15324
Sohn, R. (2022, April). AI Drug Discovery Systems Might Be Repurposed to Make Chemical Weapons, Researchers Warn. *Scientific American*. https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/
Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., Daumé, H., III, Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A. S., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., & Vassilev, A. (2023). Evaluating the Social Impact of Generative AI Systems in Systems and Society. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2306.05949
Sovacool, B. K., & Hess, D. J. (2017). Ordering theories: Typologies and conceptual frameworks for sociotechnical change. *Social Studies of Science*, *47*(5), 703–750. https://doi.org/10.1177/0306312717709363
Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, *74*(102700), 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700
Steimers, A., & Schneider, M. (2022). Sources of Risk of AI Systems. *International Journal of Environmental Research and Public Health*, *19*(6). https://doi.org/10.3390/ijerph19063641
Sun, H., Zhang, Z., Deng, J., Cheng, J., & Huang, M. (2023). Safety Assessment of Chinese Large Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2304.10436
Tan, S., Taeihagh, A., & Baxter, K. (2022). The Risks of Machine Learning Systems. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2204.09852

Teixeira, S., Rodrigues, J., Veloso, B., & Gama, J. (2022). An Exploratory Diagnosis of Artificial Intelligence Risks for a Responsible Governance. *Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance*, 25–31. https://doi.org/10.1145/3560107.3560298

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45. https://doi.org/10.1186/1471-2288-8-45

Tricco, A. C., Langlois, E. V., & Straus, S. E. (2017). *Rapid reviews to strengthen health policy and systems: a practical guide*. World Health Organization. https://ahpsr.who.int/publications/i/item/2017-08-10-rapid-reviews-to-strengthen-health-policy-and-systems-a-practical-guide

UK Department for Science, Innovation and Technology. (2023a). *A pro-innovation approach to AI regulation*. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

UK Department for Science, Innovation and Technology. (2023b). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*(2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Blili-Hamelin, B., Bollacker, K., Bomassani, R., Boston, M. F., Campos, S., Chakra, K., Chen, C., Coleman, C., Coudert, Z. D., Derczynski, L., … Vanschoren, J. (2024). Introducing v0.5 of the AI Safety Benchmark from MLCommons. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2404.12241

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). Ethical and social risks of harm from Language Models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2112.04359

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., & Isaac, W. (2023). Sociotechnical Safety Evaluation of Generative AI Systems. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2310.11986

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Wilczynski, N. L., Haynes, R. B., & Hedges Team. (2003). Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, *2003*, 719–723. https://www.ncbi.nlm.nih.gov/pubmed/14728267

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, *39*(4), 101685. https://doi.org/10.1016/j.giq.2022.101685

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, *43*(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Yampolskiy, R. V. (2016, March 29). Taxonomy of pathways to dangerous artificial intelligence. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence*. https://cdn.aaai.org/ocs/ws/ws0156/12566-57418-1-PB.pdf

Zhang, X., Chan, F. T. S., Yan, C., & Bose, I. (2022). Towards risk-aware artificial intelligence and

machine learning systems: An overview. *Decision Support Systems*, *159*(113800), 113800. https://doi.org/10.1016/j.dss.2022.113800

Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., & Huang, M. (2023). SafetyBench: Evaluating the safety of Large Language Models with multiple choice questions. In *arXiv [cs.CL]*. arXiv. https://github.com/thu-coai/SafetyBench