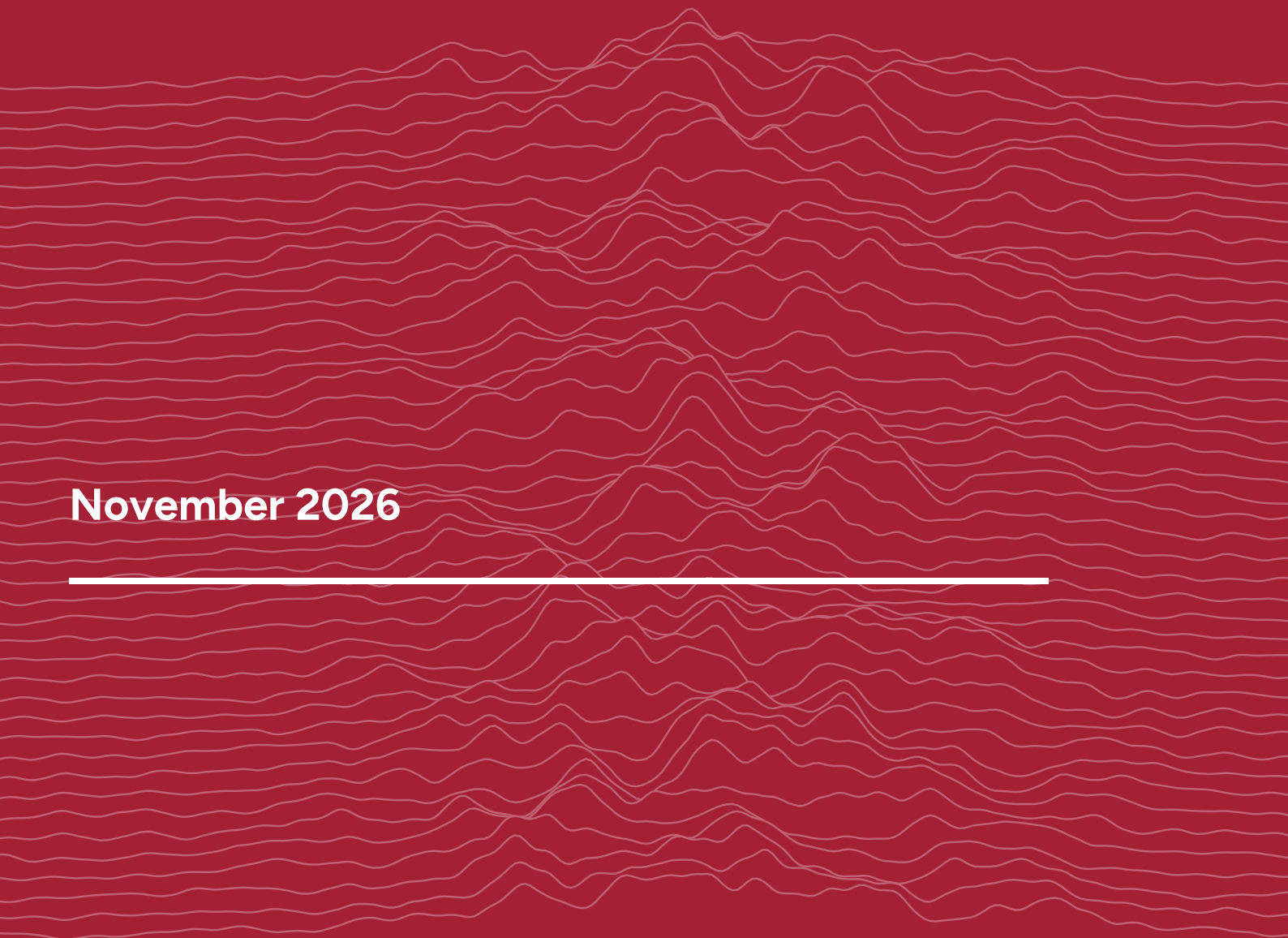
MIT AI Risk Initiative

MIT FutureTech
Innovations that Shape the World

The AI Risk Mitigation Repository

A systematic review and multi-taxonomy framework of mitigations across AI systems, organizations, and ecosystems



November 2026

The AI Risk Mitigation Repository

A systematic review and multi-taxonomy framework of mitigations across AI systems, organizations, and ecosystems

Alexander K Saeri[*,†], Kun Zhao[†], Peter Slattery, Michael Noetel, Jess Graham, Aileen Liang, Catherine Simons, Elisa Zhang, Fatimeh Al Ghannam, Soniya Agrawal, Spencer Michaels, Edla Aittokallio, Gary Ang, Keita Azuma, Isabel Barberá, Peter Barnett, Otto Barten, Samantha Augusta Bennett, Jan Brauner, Tao Burga, Ze Shen Chin, Jan Llenzl Dagohoy, Elizabeth M Daly, Tom David, Yawen Duan, Ian W Eisenberg, Rokas Gipiškis, Emily Grundy, Hans Gundlach, Adam Jones, Himanshu Joshi, Arturs Kanepajs, Noam Kolt, Sophia Lloyd George, Jayson Lynch, Nada Madkour, Claudio Mayrink Verdun, Malcolm Murray, Simon Mylius, Kwan Yee Ng, Ben Norman, Max Reddel, Graham H Ryan, Ayrton San Joaquin, Hugo Save, Daniel S Schiff, Yuanyuan Sun, Mariami Tkeshelashvili, Cristian Trout, Brian Tse, Risto Uuk, Lucia Velasco, Gabriel Weil, Neil Thompson

[*] Corresponding author: aksaeri@mit.edu; MIT FutureTech, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[†] These authors contributed equally to this work and share first authorship.

Affiliations and ORCIDs for all authors are presented at the end of the document.

Abstract

There is no shortage of proposed AI risk mitigations, but they are scattered across disciplines, described inconsistently, and difficult to compare. This lack of common structure makes it harder for researchers, developers, and policymakers to evaluate and prioritize mitigations, and to decide who should act, when, how, and against which risks. We developed a repository of AI risk mitigations, a two-axis mitigation taxonomy, and PRIMER, a reporting standard to make mitigations comparable across research and practice. Our new taxonomy separates what a mitigation aims to improve (17 domains in 3 loci: *AI System, Human & Organization, Society & Ecosystem*) from the intervention lever it uses (9 instruments in 3 functions: *Inform, Govern, Implement*). We conducted a systematic review, following a published protocol and PRISMA reporting standards, to identify 47 frameworks, taxonomies, and structured lists of AI risk mitigations from the academic and gray literature. We extracted 2,206 mitigation records and classified each by domain, instrument, actor, lifecycle stage, risk addressed, and risk management function. Mitigations concentrated on changing the AI system (55% of records) and the actions of developers (41%) and deployers (33%). But the literature also proposed mitigations to change people, institutions, and society, including *Human Oversight & Agentic Control*, *Training & Capacity-building*, and *Societal Resilience & Recovery*. As powerful AI becomes embedded across society, technical measures alone are unlikely to be sufficient. The repository, taxonomy, and PRIMER are openly available and designed to evolve with use, and we invite others to apply, test, and extend them. The next step is to establish which mitigations work and implement them to reduce the most pressing AI risks.

Keywords

AI risk mitigation, AI governance, evidence synthesis, taxonomy, classification, risk management framework

Highlights

- Systematic review of 47 documents, producing a repository of 2,206 AI risk mitigations
- Each mitigation record was classified on five taxonomies: actor, lifecycle stage, risk addressed, risk management function, and a new two-axis mitigation taxonomy
- Mitigation domain captures the mitigation focal area and instrument captures the general lever for change
- Mitigations cluster on the AI system (55%), and on developers (41%) and deployers (33%)
- PRIMER: A proposed standard for reporting mitigations consistently and comparably

Statements and declarations

Competing interests

The authors declare no competing interests.

Disclaimer

The views expressed in this paper are those of the authors and do not necessarily represent the official positions of their employers or affiliated institutions. In particular, the views expressed by Jan Brauner do not necessarily reflect those of the European Commission and are made in a personal capacity.

Author contribution statement (CRediT)

Alexander Saeri and **Kun Zhao** contributed equally and share first authorship. **Alexander Saeri**: Conceptualization, Methodology, Software, Investigation, Resources, Data curation, Writing – review & editing, Project administration. **Kun Zhao**: Methodology, Software, Investigation, Formal analysis, Writing – original draft, Visualization, Project administration. **Peter Slattery, Michael Noetel, Jess Graham:** Conceptualization, Writing – review & editing. **Aileen Liang**: Investigation, Data curation, Validation, Writing – review & editing. **Catherine Simons, Elisa Zhang, Fatimeh Al Ghannam, Soniya Agrawal, Spencer Michaels**: Investigation, Validation, Writing – review & editing. **Neil Thompson**: Supervision, Writing – review & editing. **All other authors:** Methodology, Investigation, Writing – review & editing.

Data and code availability

Data and code required to reproduce analyses in this manuscript are publicly available at https://osf.io/ft6mg/

AI use

Large language models were used to assist in the screening, extraction, and classification stages of the systematic review, subject to human review. These processes are described in the Methods section. Large language models were used to assist with code for data analysis and data visualizations, as well as with manuscript editing. All content was reviewed and verified by the authors, who take full responsibility for the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The MIT AI Risk Initiative and MIT FutureTech are funded by a range of organizations; none of those organizations influenced the design, collection, analysis, interpretation, or reporting of the data.

Acknowledgments

We gratefully acknowledge the contributions of Ben Bucknall, Stephen Casper, Jiyeon Cho, Gwyn Glasser, Koen Holtman, Michael J Howell, Anastassia Kornilova, Reva Schwartz, and expert panel members who preferred not to be named.

Contents

Introduction 7
- There are many proposed solutions for AI risks 7
- The need for a structured specification of AI risk mitigations 8
- The current work 9

Results 12
- The two-axis AI risk mitigation taxonomy 15
- Distribution of mitigations by domains and instruments in the AI Risk Mitigation Taxonomy 19
- Distribution of mitigations by AI actor, AI lifecycle stage, AI risk addressed, and AI risk management function 24
- AI risk mitigations across multiple dimensions: *Who* does *what* and *when* to address *which* risks? 27

Discussion 34
- Overview of findings 34
- Insights from the literature: What clusters and gaps across taxonomies reveal about mitigations 34
- Classifying AI risk mitigations: Learnings and challenges 36
- A proposed mitigation reporting standard 37
- Practical implications and future directions 41
- Limitations 42
- Conclusion 43

Methods 44
- Overview of approach 44
- Systematic search of peer-reviewed and gray literature 45
- Document screening 46
- Included documents 46
- Extraction 47
- Novel AI risk mitigation taxonomy development 48
- LLM-assisted classification and validation 50

References 52

Author affiliations and ORCIDs 60

Supplementary tables 62

List of tables

Table 1. AI Risk Mitigation Taxonomy: Domain axis 17
Table 2. AI Risk Mitigation Taxonomy: Instrument axis 19
Table 3. Distribution and robustness of mitigation domains 20
Table 4. Distribution and robustness of mitigation instruments 22
Table 5. AI actor categories in the AI Risk Mitigation Repository 24
Table 6. AI lifecycle stage categories in the AI Risk Mitigation Repository 25
Table 7. AI risk categories in the AI Risk Mitigation Repository 26
Table 8. NIST AI Risk Management Framework function categories in the AI Risk Mitigation Repository 27
Table 9. Characteristics of included documents 48

List of figures

Figure 1. Overview of methodology 13
Figure 2. An example mitigation record from source document to classification 14
Figure 3. AI Risk Mitigation Taxonomy 16
Figure 4. Distribution of mitigation records across mitigation domains 21
Figure 5. Distribution of mitigation records across mitigation instruments 23
Figure 6. Distribution of mitigation records by mitigation domain and instrument 28
Figure 7. Distribution of mitigation records by mitigation domain and instrument for the three largest domains of risk addressed 29
Figure 8. Most frequent AI actor–risk domain–mitigation domain combinations in the AI Risk Mitigation Repository 31
Figure 9. Distribution of mitigation records by AI lifecycle stage and mitigation domain 32
Figure 10. Distribution of mitigation records by AI lifecycle stage and mitigation instrument 33
Figure 11. Expert-rated probability of severe or catastrophic harm against mitigation record count, by AI risk subdomain 35
Figure 12. PRIMER: Preferred Reporting Items for Mitigations of Emerging Risks 39
Figure 13. PRIMER worked example 40
Figure 14. PRISMA study flow diagram 44

Introduction

What mitigations exist to address artificial intelligence (AI)[1] risks? Rapid increases in AI capabilities and widespread adoption of AI are exposing society to many different risks from AI, both familiar and novel, some of which may lead to enduring societal or even catastrophic harm (Bengio et al., 2026). Common language and frameworks are emerging to organize these risks (Slattery et al., 2026) and capture incidents of harm (OECD.AI Policy Observatory, 2026b; Responsible AI Collaborative, 2026) in order to better guide mitigations. Yet, as observers have cautioned, the AI risk mitigation landscape remains fragmented, confounding progress on timely and effective risk management. Both the International AI Safety Report (Bengio et al., 2026) and the United Nations' Preliminary Report of the Independent International Scientific Panel on AI (United Nations, 2026) have noted the current fragmented regulatory landscape and the uneven distribution of governance instruments. They have called for shared evidence standards and global coordination in response. To that end, we performed a systematic review of research on AI risk mitigations and identified >2,000 mitigations across 47 documents. We classified these mitigations using four established taxonomies and one new taxonomy, and we propose a new reporting tool, PRIMER (Preferred Reporting Items for Mitigations of Emerging Risks), to help researchers and practitioners better communicate and organize the current state of knowledge on AI risk mitigations.

There are many proposed solutions for AI risks

There has been a proliferation of proposed safeguards, governance mechanisms, and best practices for addressing risks from AI (Crichton et al., 2025; OECD.AI Policy Observatory, 2026c; Uuk et al., 2024). Few offer systematic guidance for defining, comparing, and prioritizing mitigations. The OECD (Organisation for Economic Co-operation and Development) lists over 1,000 tools and metrics for trustworthy AI, including certification and standards, change management processes, and educational materials (OECD.AI Policy Observatory, 2026a), alongside nearly 2,500 national AI policy initiatives (OECD.AI Policy Observatory, 2026c). Technical and organizational standards form another layer, with the AI Standards Hub cataloging more than 500 AI-related standards from bodies including the International Organization for Standardization (ISO), the International Electrotechnical Commission (IEC), the Institute of Electrical and Electronics Engineers (IEEE), and the International Telecommunication Union (ITU) (Alan Turing Institute, 2026; Schiff, 2026). These standards carry normative and informative references to further standards, creating cascading chains of requirements that are difficult to understand and follow (Schiff, 2026). Regulation for AI is being constantly revised and varies widely across jurisdictions, making it hard for even motivated organizations to comply. The EU AI Act sets layered requirements for high-risk AI systems, from risk management systems and data governance through to human oversight and cybersecurity (European Parliament and Council of

[1] While there are many definitions, we refer to AI broadly as "machine-based models or systems capable of performing tasks that typically require human intelligence, such as generating text" (Bengio et al., 2026, p. 147).

the European Union, 2024), and its General-Purpose AI Code of Practice specifies safety and security measures for providers of models with systemic risk (European Commission, 2025).

This abundance of proposed policy and practice has become its own challenge (Ramchurn et al., 2025). The same name may refer to several practices; for example, "red teaming" covers several methods, addresses different threat models, and carries little consensus on who performs it (Feffer et al., 2024). The concept of "transparency" could mean making a model's reasoning legible (i.e., improving model interpretability; see Larsson & Heintz, 2020), disclosing information to regulators and users (i.e., improving organizational accountability), or public reporting on practices (i.e., improving knowledge sharing). "Risk management" appears in most AI risk frameworks but is defined inconsistently and varies in scope and specificity (Saeri et al., 2025; Schuett et al., 2023). This lack of common language can hamper coordination between actors to address pressing risks from AI, or allow actors to undermine compliance through "safety washing" and other strategic framing that promotes or demotes favored risks.

The emerging AI risk mitigations literature draws on different practice backgrounds and is fragmented by discipline. Some mitigations are borrowed from cybersecurity and safety-critical industries (e.g., incident reporting from aviation and healthcare), or imported from other regulatory domains (e.g., third-party audits from financial, environmental, and health regulation). Others are unique to AI (e.g., alignment and interpretability methods). Without clear standardization in terminology and scope, practitioners bear the burden of reconciling overlapping obligations (Tanner et al., 2026). Finally, responsibility is dispersed, characterized by a "many hands" problem in accountability and roles across a range of technical and social domains (Schiff et al., 2021).

The need for a structured specification of AI risk mitigations

Even when mitigations are named consistently, they are rarely specified in enough detail to act on: they lack consistent operationalization. When advocates call for stronger action on risks such as AI-enabled cyberattacks, the actions that this could translate to are wide-ranging and leave open who should address it (the developer, an independent third party, a government body, end users), when (before release or throughout deployment), and whether the solution involves a technical innovation, a voluntary practice, an industry standard, or a legal requirement. The discipline of implementation science emphasizes that naming strategies (e.g., "test models before release") is not sufficient; unless the strategies are specified by actor, action, target, and timing, they cannot be replicated, evaluated, or held to account (Proctor et al., 2013). The same gap is visible in AI: When organizations reported their risk management measures in the first cycle of the G7 Hiroshima AI Process (HAIP) Reporting Framework (OECD, 2025), responses were largely high-level and difficult to verify, and variability between submissions limited their comparability (Bogen et al., 2026; Perset & Fialho Esposito, 2025).

Together, these challenges suggest that a significant barrier to addressing the most pressing AI risks is not a lack of promising solutions. Instead, a barrier is inadequate shared infrastructure for defining and operationalizing them: guiding *who* should act, *when* and *where* mitigations should be carried out, *what* problems should be addressed, *what* methods should be used, and how all of these should be prioritized (see also Schiff, 2026).

Shared taxonomies are a precondition for operationalizing and coordinating action on AI risks (US-EU Trade and Technology Council, 2022). Nuclear safety, pandemic preparedness, and cybersecurity offer precedents for establishing common knowledge through standardized classification vocabularies (Mengesha, 2026). Efforts to organize the sprawl of AI risk mitigations are under way. Earlier approaches aimed to translate AI ethics principles into a typology of practices that could be mapped to the machine learning development pipeline (Morley et al., 2019). The National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF; NIST, 2023), a widely adopted structuring system, has been used for classifying risk management practices against its four functions and across the lifecycle, for example, among developers of foundation models (Barrett et al., 2025). While these overlap with mitigations, they are broad functions rather than specific implementable actions themselves. The OECD's risk management framework (OECD, 2023) offers a systematic way to govern and manage risks across the AI system lifecycle, organizing examples of treatments by risks addressed and technical versus process approaches. Lu et al. (2024) operationalized responsible AI guidance into a catalog of reusable design patterns, spanning governance (e.g., industry, organizational, team), process, and product angles. Crichton et al. (2025) distilled more than 7,000 recommendations across 52 guidance documents into 258 practices, scoped to organizations and excluding regulation and legislation by design. Maas (2023) paired key actors with levers of governance across the advanced AI governance literature. Fewer examples span all of these boundaries, covering technical, organizational, and societal mitigations, though this is the level at which AI risk is increasingly discussed (United Nations, 2026).

The current work

We build on these recent efforts by conducting a systematic review of AI risk mitigations. We included documents that described frameworks, taxonomies, or other structured lists of AI risk mitigations.

Defining AI risk mitigations

We adopt an intentionally broad definition of *AI risk mitigations* in this review: "an action that reduces the likelihood or impact of an unfortunate occurrence that may emerge from the development, deployment, or use of systems or machines capable of performing tasks that typically require human intelligence" (after Actuarial Standards Board, 2014; Aven et al., 2018; Bengio et al., 2026). We include mitigations that do so indirectly by enabling risks to be identified, assessed, monitored, or governed.

We recognize that our definition of *risk mitigations* overlaps and is in tension with other related concepts in the risk management and AI safety literatures (Bengio et al., 2026; ISO, 2018, 2022). In its narrowest sense, a mitigation only reduces the severity of consequences after a harm-causing incident, as on the right side of a bow-tie diagram, in explicit contrast to preventive measures (Abdelwanis et al., 2024). In risk management standards, risk mitigation is a type of *risk treatment* that reduces negative consequences, as distinct from other risk treatments (e.g., that reduce likelihood, share the risk with other parties, avoid the risk; ISO 31073; ISO, 2022, Section 3.3.32). Risk treatment is one step in a *risk management process* defined as "the systematic application of policies, procedures and practices to the activities of communicating and consulting, establishing the context and assessing, treating, monitoring, reviewing, recording and

reporting risk" (ISO 31000; ISO, 2018, Section 6.1). The International AI Safety Report and the NIST AI Risk Management Framework describe comparable four-part structures for AI risk management: governing, identifying, evaluating, and mitigating risk (Bengio et al., 2026), and *Govern*, *Map*, *Measure*, and *Manage* (NIST, 2023). The International AI Safety Report defines *risk mitigation* as "the process of prioritising, evaluating, and implementing controls and countermeasures to reduce identified risks" (p. 110), but its examples "access controls, continuous monitoring, and if-then commitments" span multiple steps in a traditional risk management process (Bengio et al., 2026).

In the AI safety and AI governance literatures the term is used more broadly still. Frontier safety frameworks use "mitigations" for security and deployment measures, including monitoring (Anthropic, 2026; Google DeepMind, 2026). Catalogs of risk management "measures" (Gipiškis et al., 2024) or "controls" (Eisenberg et al., 2025; ISO/IEC 42001; ISO/IEC, 2023, Annex A) extend further to evaluation, governance, and disclosure activities. We adopt the broadest usage for the purpose of inclusion in this systematic review. We did so because the repository and taxonomy are intended as a common framework for all actors seeking to address risks from AI, and this work spans a vast range of related activities. While the functions of risk management are separable in principle, in practice they may be difficult to disentangle, and individual measures are rarely effective alone, instead requiring a portfolio of related approaches (Uuk et al., 2024, also see Discussion section).

Research aims

The study had three aims:

1. To identify what mitigations have been proposed in recent frameworks, taxonomies, and structured lists.
2. To organize these mitigations into taxonomies to support standardization and implementation. We developed a new high-level taxonomy of mitigations to support this aim.
3. To use this classification to answer questions about the AI risk mitigation landscape: which mitigations are the most and least commonly reported, which AI risks are addressed, and which actors, lifecycle stages, and risk management functions are implicated.

Following a protocol and PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) reporting standards (Page et al., 2021), we searched, screened, and extracted from 47 documents published between 2018 and 2026, producing over 2,000 mitigation records (see Methods for details of scope and screening decisions). Each record was classified against five taxonomies: four established ones (actor, lifecycle stage, risk addressed, and risk management function) and a new two-axis taxonomy of mitigations. The new taxonomy was developed to meet best-practice criteria, including structure and content that is concise, robust, comprehensive, extendible, and explanatory, as well as mutually exclusive and collectively exhaustive on each dimension (Nickerson et al., 2013). This taxonomy was validated and refined through expert consultation, so that categories reflected practitioner understanding and usability in addition to fidelity to the source literature. We consider this a critical requirement for the taxonomy to support communication about mitigations across research and practice areas, given that coordination at scale is one of the most pressing AI governance challenges (Bengio et al., 2026; United Nations, 2026). Classification was assisted by large language models (LLMs), each

tested against independent human raters to confirm LLM-human agreement was at least as high as human-human agreement.

Research contributions

This work makes four contributions:

1. We provide a comprehensive and systematic account of the AI risk mitigation landscape. Systematic reviews have focused on AI governance and technical methods (Shi et al., 2024; Sigfrids et al., 2022; Wirtz et al., 2022) and previous syntheses have been largely narrative or purposive in scope. Our own preliminary evidence scan of 13 documents established feasibility but was weighted toward frontier AI and developer organizations (Saeri et al., 2025).
2. We introduce a novel two-axis mitigation taxonomy that separates a mitigation's focal area (or domain) from the lever through which it acts (or instrument). Existing single-axis classification schemes often required combining distinct organizing principles – what a mitigation targeted, what risk it addressed, when it should take place, and through what mechanism – into one set of categories. Crossing the two axes holds key mitigation concepts together on one dimension while accounting for their variation on the other, following intervention taxonomies from other disciplines (Haddon, 1972; Hood, 1983; Michie et al., 2011).
3. We provide a multi-dimensional classification of every mitigation record, following implementation science guidance on specifying measures by actor, action, target, and timing (Proctor et al., 2013).
    a. The AI actor taxonomy (adapted from Australian Government, 2024; NIST, 2023) asks *who* is responsible, addressing the many-hands problem of dispersed accountability.
    b. The AI lifecycle taxonomy (adapted from NIST, 2023; OECD, 2022) asks *when*, locating where in development or operation a mitigation must be implemented.
    c. The AI risk taxonomy (Slattery et al., 2026) asks *what* risk a mitigation is designed to act on, making explicit the link between problem and solution.
    d. The AI risk management function taxonomy (NIST, 2023; OECD, 2022) aligns mitigations with a framework already used to organize risk management activity.
    e. The new two-axis AI Risk Mitigation Taxonomy asks what a mitigation aims to change and through what methods.
4. We provide an empirical picture of crowded and neglected areas in AI risk mitigation research. The repository and taxonomy are published openly so that others can extend the classification and assess the effectiveness and feasibility of mitigations in addressing the most pressing AI risks.

Results

Figure 1 provides an overview of our systematic review and taxonomy construction methodology. These are described in detail in the Methods section. We identified documents with AI risk mitigations through (1) a systematic search of academic and gray literature, and (2) expert consultation. We screened documents for relevance and extracted and classified individual mitigations using a combined human and LLM process. In total, we identified 47 relevant documents and extracted and classified 2,206 mitigation records against five taxonomies: a new two-axis taxonomy of AI risk mitigations that was developed as part of this work, and four established taxonomies. Figure 2 presents an example of how a mitigation record from a relevant document was extracted and classified. We organize the results in three sections:

1. We describe a novel **AI risk mitigation taxonomy** with two axes: mitigation *domains* (what a mitigation aims to improve) and mitigation *instruments* (the lever through which it acts).
2. We describe distributions of individual mitigations within the two axes of the **AI Risk Mitigation Taxonomy**, as well as each of the remaining four taxonomies: **AI actor**, **AI lifecycle stage, AI risk addressed**, and **AI risk management function**, based on the systematic review.
3. We describe patterns *across* all five taxonomies above, to explore "*Who* does *what* and *when* to address *which risks*?"

Figure 1. Overview of methodology

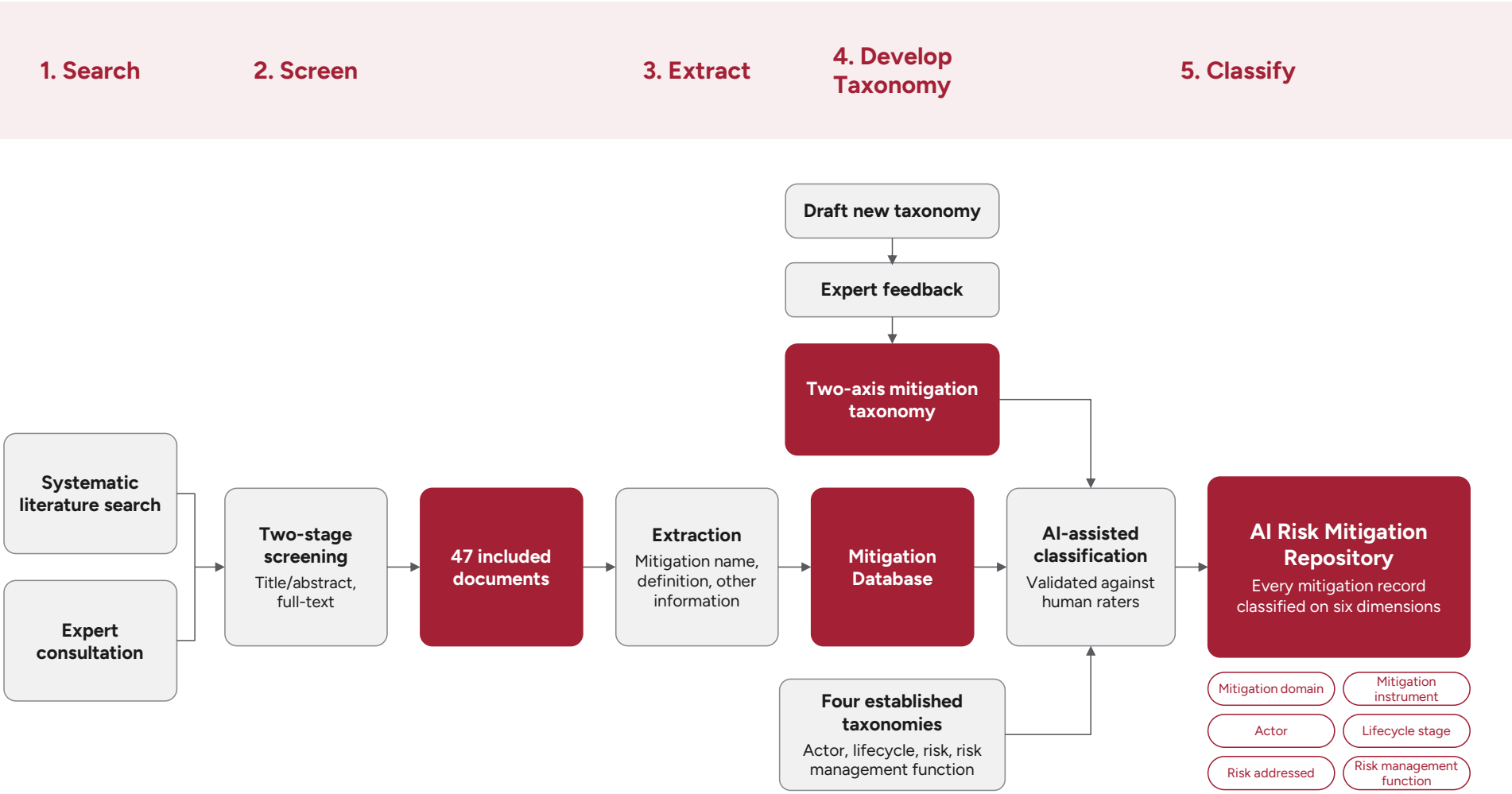


Figure 2. An example mitigation record from source document to classification

1. Source document

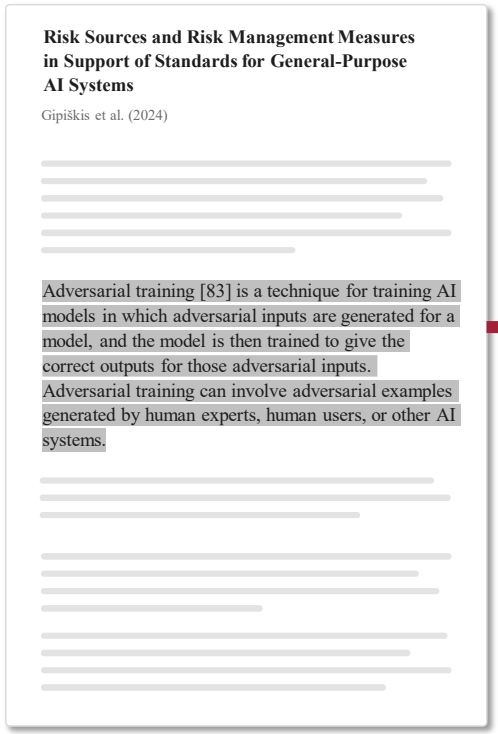
A passage describing a mitigation is located in the document.

2. Extracted mitigation record

The mitigation is recorded as described by the document author: name, definition, additional information, and source.

3. Classified against five taxonomies

The mitigation record is classified based on mitigation name, definition, additional information, document context



Extract

MITIGATION NAME
Adversarial training

DEFINITION
"Adversarial training [83] is a technique for training AI models in which adversarial inputs are generated for a model, and the model is then trained to give the correct outputs for those adversarial inputs. Adversarial training can involve adversarial examples generated by human experts, human users, or other AI systems."

ADDITIONAL INFORMATION
— (none provided)

SOURCE
Gipiskis et al. (2024)

RECORD ID
`A2213_Gipiskis2024`

Classify

A · AI RISK MITIGATION TAXONOMY (TWO AXES)

MITIGATION DOMAIN
AI System: Robustness & Adversarial Resistance
Changes the AI model itself, allowing it to work correctly when faced with adversarial inputs designed to fool it.

MITIGATION INSTRUMENT
Implement: Technical Artifacts
Protection is built into the model through training, rather than a procedure that people carry out.

B · EXISTING TAXONOMIES

AI ACTOR
Developer
Training a model is done by the organization that designs and builds the AI system.

AI LIFECYCLE STAGE
Build & Use Model
Applied while the model is being trained, before deployment.

NIST AI RMF FUNCTION
Manage
Acts directly to reduce an identified risk, rather than governing, mapping or measuring it.

AI RISK ADDRESSED
Privacy & Security: AI System Security Vulnerabilities & Attacks
Defends against attacks that exploit the model's weaknesses to cause unsafe or incorrect outputs.

The two-axis AI risk mitigation taxonomy

We developed a novel taxonomy of AI risk mitigations with two axes: (1) mitigation *domain*, what the mitigation aims to change or improve, clustered into three higher-level loci (*AI System, Human & Organization, Society & Ecosystem*); and (2) mitigation *instrument*, the mechanism through which change is produced, clustered into three higher-level functions (*Inform, Govern, Implement*).

The development process is reported in detail in Methods. We initially intended to adapt an existing framework from the systematic review as a "best-fit taxonomy" on which to classify all other mitigation records, but the included documents were too diverse for any single framework to serve this role, and existing frameworks were either too high-level, too narrow, or conflated multiple organizing principles. We then constructed and iterated on a novel AI risk mitigation taxonomy through inductive-deductive iteration on the extracted mitigation records, comparison with other existing taxonomies, and feedback from 45 experts.[2] The novel AI Risk Mitigation Taxonomy was designed to accommodate distinct organizing principles by separating where change occurs (domain) from the lever through which change is produced (instrument). This had the advantage of presenting clusters of mitigations together (e.g., the many "red teaming" mitigation records) on one axis while distinguishing them on the other.[3]

We expect the taxonomy to evolve as the mitigation landscape develops and as the taxonomy is applied to corpora beyond the review literature. The AI Risk Mitigation Taxonomy is summarized in Figure 3, with definitions and examples provided in Table 1 (domain axis) and Table 2 (instrument axis).

[2] The classification challenges that shaped these decisions (such as the breadth of what counts as a mitigation, uneven specificity across records, and the blurred line between core interventions, supporting techniques, and enabling conditions) are described in the Discussion.

[3] An example is "whistleblower protections", a cluster of mitigations that encourage and support reporting of critical AI safety risks. Almost all of these records were classified under a single mitigation domain (*Integrity & Ethics*), but were separated across several mitigation instruments (e.g., *Law & Regulation* to protect whistleblowers from retaliation; *Internal Policies* where companies describe how insiders can report problems; and *Other* instruments such as the creation of an organization that can receive reports from whistleblowers).

Figure 3. AI Risk Mitigation Taxonomy

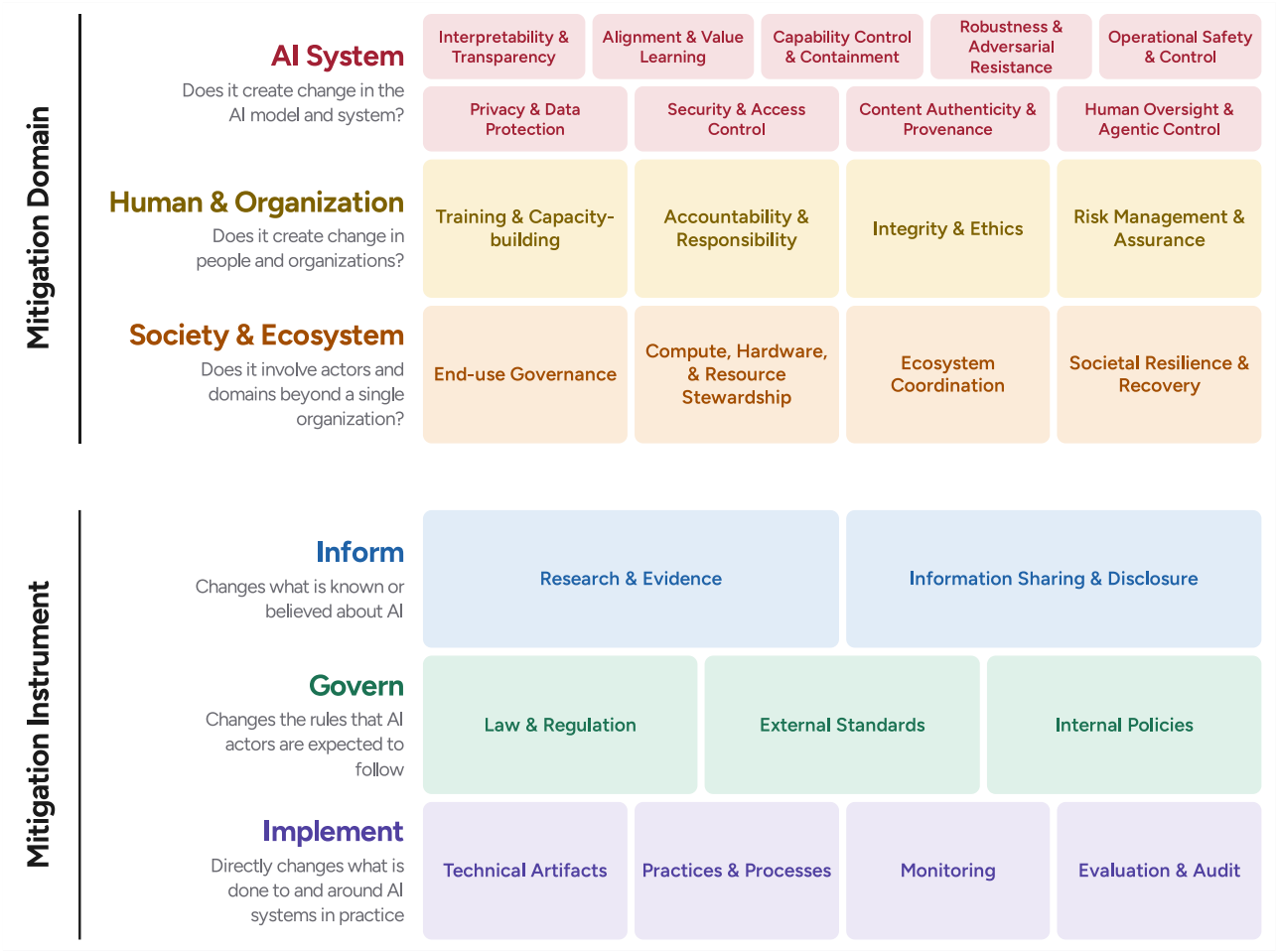


Table 1. AI Risk Mitigation Taxonomy: Domain axis

Locus / Domain	Short label	Definition	Examples
AI System	***AI System***	***Mitigations at the level of the AI system that target the properties, behavior, and capabilities of the model and system to allow it to be developed, deployed, and operated safely and as intended.***	
Interpretability & Transparency	Interpretability	Mitigations that make a model's behavior and internals legible.	• Model cards • SHAP methods • Linear probes
Alignment & Value Learning	Alignment	Mitigations that shape a model's goals and values. Includes reducing bias, unfairness, and discrimination.	• RLHF • Constitutional AI • Debiasing/detoxification via training
Capability Control & Containment	Capability	Mitigations that bound, measure, or halt what the model can do.	• Dangerous-capability evaluations • Unlearning • Staged release
Robustness & Adversarial Resistance	Robustness	Mitigations that improve AI model reliability under stress/attack.	• Adversarial training • Red-teaming • Data-poisoning defenses
Operational Safety & Control	Operational	Mitigations that keep a deployed system's live behavior safe and within intended bounds during operation. Includes gating (filter/block unsafe inputs/outputs), confining (bound what the running system can access or do), watching (monitor behavior for anomalies), and instructing (steer behavior via inference-time instructions).	• I/O content filters and guardrails • Sandboxing • Real-time monitoring
Privacy & Data Protection	Privacy	Mitigations that protect personal or sensitive data.	• PII removal and anonymization • Federated learning • Data minimization
Security & Access Control	Security	Mitigations that prevent unauthorized access, availability, and abuse. Includes responding to or recovering from incidents.	• Model-weight protection • Cybersecurity controls for AI systems • Incident response plans
Content Authenticity & Provenance	Authenticity	Mitigations that prove content origin and counter synthetic deceptive content.	• Watermarking outputs • C2PA provenance tracking • AI content labeling
Human Oversight & Agentic Control	Oversight	Mitigations that allow humans to oversee/intervene, including bounding agent autonomy.	• HITL review interfaces • Kill-switch, interruptibility • Agent activity logs
Human & Organization	***Organization***	***Mitigations at the organizational level that target the capabilities, actions, and motivations of humans to allow the responsible development, oversight, and use of AI.***	
Training & Capacity-building	Training	Mitigations that build the AI-safety knowledge, skills, and expertise of people who develop, deploy, or work with AI so they can do this work safely and build safer systems.	• Red-team / evaluator training • Certification for AI-safety practitioners • Internal staff safety and ethics training
Accountability & Responsibility	Accountability	Mitigations consisting of rules, roles, internal structures, and external oversight through which an organization and its members are held to account to engage in responsible conduct concerning AI. Includes disclosure and reporting for transparency.	• Board oversight • Incident reporting and disclosure • Third-party audit
Integrity & Ethics	Ethics	Mitigations that provide protections and disclosure mechanisms through which an organization's AI-related conduct can be scrutinized for integrity.	• Whistleblower protection channels • Independent ethics review boards • Conflict of interest disclosure
Risk Management & Assurance	Risk Management	Mitigations that identify, aggregate, and gather evidence that an organization's AI risks are understood and controlled to an acceptable level.	• Risk registers and risk matrices • Safety / assurance cases • Risk tolerance statements

Locus / Domain	Short label	Definition	Examples
Society & Ecosystem	***Ecosystem***	***Mitigations at the level of the wider ecosystem that target the shared rules, institutions, and conditions to allow AI to be governed, contested, and absorbed safely across society.***	
End-use Governance	End-use	Mitigations that govern how AI is used safely by individuals or organizations in the end-user role, across both organizations and the wider ecosystem. Includes both user protections and user conduct.	• Acceptable-use policies • Individual practices such as output cross-checking • User rights: opt-out, contestability
Compute, Hardware, & Resource Stewardship	Compute	Mitigations that ensure the compute, hardware, and physical resources underpinning AI are controlled, secured, and sustainably managed by both organizations and across the eco-system.	• Compute-reporting thresholds • Hardware/chip export controls • Renewable-energy procurement
Ecosystem Coordination	Coordination	Mitigations that enable alignment and coordination across organizations/sectors/nations to govern AI safely. Includes shared safety infrastructure and resources. Includes market mechanisms, such as mitigations that change cost, rewards, and incentives at the market level (e.g., pricing, subsidies, funding, tax).	• Regulatory and oversight bodies • International agreements • Liability regimes
Societal Resilience & Recovery	Resilience	Mitigations that enable society to prepare for, withstand, and recover from the consequences of AI.	• Public AI literacy campaigns • Societal incident registries • Crisis-management plans for AI catastrophes

Note. Taxonomy axis also includes "Other" categories for each locus, and one top-level "Other" category for mitigations that were too broad for any single category, cross-cutting, or otherwise unclassifiable.

Table 2. AI Risk Mitigation Taxonomy: Instrument axis

Function / Instrument	Short label	Definition	Examples
Inform	***Inform***	***Mitigations that work by changing what is known or believed about AI.***	
Research & Evidence	Research	Mitigations involving the generation and dissemination of new knowledge and evidence.	• Research studies • Research datasets • Foresight scanning
Information Sharing & Disclosure	Disclosure	Mitigations involving the open communication and sharing of information across actors.	• Disclosure and reporting of incidents • Information sharing on industry registries
Govern	***Govern***	***Mitigations that work by changing the rules that AI actors are expected to follow.***	
Law & Regulation	Laws	Mitigations involving the creation or adoption of systematic rules by an external authority with jurisdiction over the target.	• EU AI Act • Export controls on AI chips • Statutory requirement to conduct risk assessments
External Standards	Standards	Mitigations involving the creation and adoption of shared systematic rules from a peer/industry body.	• NIST AI RMF • OECD AI principles • Industry safety pledges
Internal Policies	Policies	Mitigations involving the creation and adoption of systematic rules by an actor for itself (and internally for its staff).	• Internal responsible AI policies • A named safety framework an organization authors
Implement	***Implement***	***Mitigations that work by directly changing what is done to and around AI systems in practice.***	
Technical Artifacts	Technical	Mitigations that are built into the AI system or its supporting infrastructure.	• Content filters • Sandboxing • Watermarking
Practices & Processes	Practices	Mitigations consisting of procedures, workflows, or ways of working that humans carry out.	• Incident-response procedures • Review and sign-off workflows • Pre-deployment checklists
Monitoring	Monitoring	Mitigations consisting of the continuous observation of an AI system, its behavior, use, or effects to surface problems on an ongoing basis.	• Run-time behavior monitoring • Logging/audit trails • Drift detection
Evaluation & Audit	Evaluation	Mitigations that primarily involve testing or checking against a standard or expectation.	• Testing a model against benchmarks • Third-party audits • Dangerous-capability evaluations

Note. Taxonomy axis includes an "Other" category for mitigations that were too broad for any single category, cross-cutting, or otherwise unclassifiable.

Distribution of mitigations by domains and instruments in the AI Risk Mitigation Taxonomy

Domain axis: What do mitigations aim to change or improve?

We structured the domain axis of the AI Risk Mitigation Taxonomy into 3 loci (top-level categories) and 17 domains (subcategories). The three loci were *AI System*, *Human & Organization*, and *Society & Ecosystem*, which were used to organize and cluster mitigation domains. More than half (55%) of reported mitigations aimed to address risk by changing the *AI system*; more than a quarter (28%) aimed to change *Human & Organization*; 15% aimed to change *Society & Ecosystem*; and 1.1% were classified as *Other*.

The most commonly reported mitigation domains were *Accountability & Responsibility* (13% of mitigation records), *Security & Access Control* (13%), and *Risk Management & Assurance* (12%). The least commonly reported domains were *Integrity & Ethics* (0.7%), *Training & Capacity-building* (1.3%), and *Compute, Hardware, & Resource Stewardship* (1.3%). Figure 4 presents the

distribution of mitigation domains across the database; Table 3 reports detailed statistics on the same.

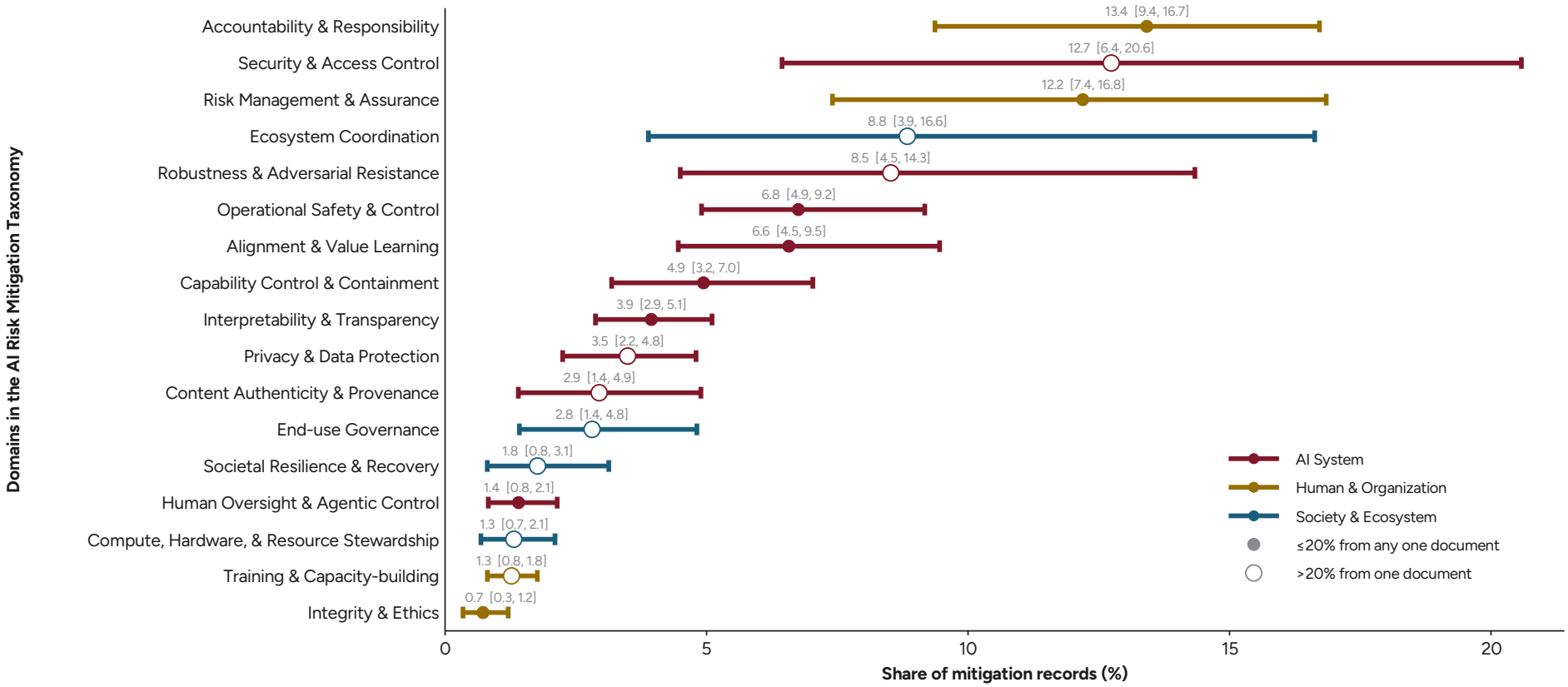
We conducted a robustness analysis to investigate whether the patterns of mitigations across domains were being driven by specific documents. For example, a document that describes many different types of interpretability techniques could have artificially inflated the share of all mitigations classified as *Interpretability & Transparency*. Table 3 reports the largest single source document for each mitigation domain (e.g., NIST, 2024, provided 16% of all mitigation records classified as *Interpretability & Transparency*), and how the domain's rank (i.e., how large or small a share of all mitigations it represented) varied if any single document was deleted. Figure 4 presents the confidence interval for each mitigation domain according to the same checks. Most mitigation domains were stable under the robustness checks. For example, *Accountability & Responsibility* held its rank as the most or second-most reported mitigation domain (13.4%, 95% CI [9.4, 16.7]) under every single-document deletion. *Ecosystem Coordination,* along with *Security & Access Control,* were the least robust mitigation domains. In the former case, its interval was the widest relative to its own estimate (8.8%, 95% CI [3.9, 16.6]). Its rank moved between 3rd and 7th across single-document deletions, and 38% of its records came from a single source (Maas, 2023).

Table 3. Distribution and robustness of mitigation domains

Locus / Domain	Documents	Largest single source	Rank across deletions	n	%
AI System	***46***			***1,223***	***55.4***
Interpretability & Transparency	27	NIST2024 (16%)	9–9	87	3.9
Alignment & Value Learning	32	Wang2025 (14%)	6–7	145	6.6
Capability Control & Containment	30	DSIT2023 (15%)	8–8	109	4.9
Robustness & Adversarial Resistance	34	Ma2025 (26%)	4–5	188	8.5
Operational Safety & Control	35	Ma2025 (9%)	5–7	149	6.8
Privacy & Data Protection	23	Crichton2025 (21%)	10–11	77	3.5
Security & Access Control	37	Crichton2025 (39%)	1–5	281	12.7
Content Authenticity & Provenance	19	NIST2024 (37%)	10–12	65	2.9
Human Oversight & Agentic Control	17	NIST2024 (13%)	14–15	31	1.4
AI System (Other)	*30*			*91*	*4.1*
Human & Organization	***36***			***625***	***28.3***
Training & Capacity-building	18	NIST2024 (21%)	15–16	28	1.3
Accountability & Responsibility	34	Crichton2025 (18%)	1–2	296	13.4
Integrity & Ethics	11	Maas2023 (19%)	17–17	16	0.7
Risk Management & Assurance	26	NIST2024 (20%)	2–3	269	12.2
Human & Organization (Other)	*7*			*16*	*0.7*
Society & Ecosystem	***34***			***334***	***15.1***
End-use Governance	18	NTC2602024 (27%)	11–12	62	2.8
Compute, Hardware, & Resource Stewardship	14	Maas2023 (28%)	14–16	29	1.3
Ecosystem Coordination	25	Maas2023 (38%)	3–7	195	8.8
Societal Resilience & Recovery	15	Bengio2026 (21%)	13–13	39	1.8
Society & Ecosystem (Other)	*5*			*9*	*0.4*
Other	*16*			*24*	*1.1*
Total	47			2,206	100.0

Note. N = 2,206 mitigation records, from a total of 47 source documents. "Documents" is the number of source documents contributing at least one record to that category. For this reason, rows in the "Documents" column sum to greater than 47. Rank across deletions is the range of ranks the category took when each document in turn was removed. The largest single source and its share of the category's records indicate single-document dependence.

Figure 4. Distribution of mitigation records across mitigation domains



Note. Intervals are 95% cluster bootstrap intervals over source documents (20,000 resamples). A hollow marker indicates that more than 20% of the category's records came from a single document, so its position in the ranking is less stable. Shares are of all eligible mitigation records, with the remainder consisting of "Other" or cross-cutting codes. N = 2,206 eligible mitigation records from 47 documents.

Instrument axis: What are the levers through which mitigations act?

We structured the instrument axis of the AI Risk Mitigation Taxonomy into 3 functions (top-level categories) and 9 instruments (subcategories). The three functions were *Inform, Govern*, and *Implement*, which were used to organize and cluster mitigation instruments. Nearly three-quarters (71%) of reported mitigations changed what is done to and around AI systems in practice (*Implement*). Fewer mitigations changed rules that AI actors are expected to follow (12%, *Govern*), or changed what is known or believed about AI (8.5%, *Inform*). A further 7.8% were classified as *Other*.

The most reported mitigation instruments were *Technical Artifacts* (26% of records), followed by *Practices & Processes* (24%), and *Evaluation & Audit* (17%). The least reported instruments were *Monitoring* (4.9%), *Internal Policies* (4.9%), *Research & Evidence* (2.6%), and *External Standards* (2.3%). Figure 5 presents the distribution of mitigation instruments across the database; Table 4 reports detailed statistics on the same.

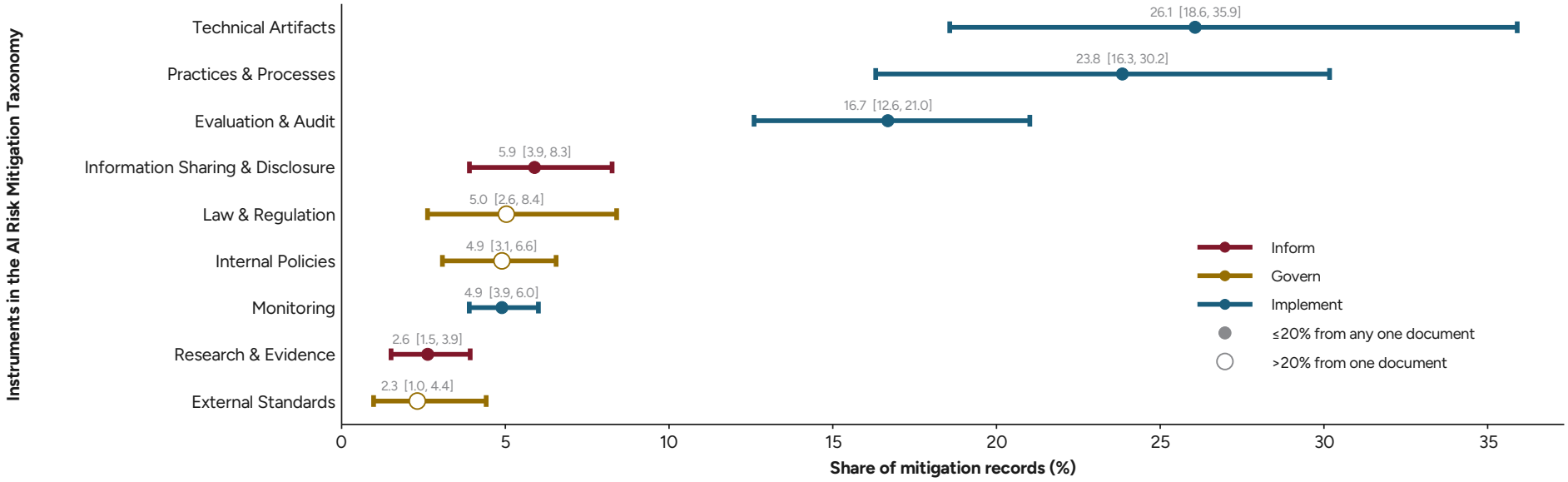
Table 4. Distribution and robustness of mitigation instruments

Function / Instrument	Documents	Largest single source	Rank across deletions	n	%
Inform	***31***			***188***	***8.5***
Research & Evidence	22	DSIT2023 (16%)	8–9	58	2.6
Information Sharing & Disclosure	25	DSIT2023 (17%)	4–5	130	5.9
Govern	***35***			***270***	***12.2***
Law & Regulation	23	Maas2023 (28%)	4–7	111	5.0
External Standards	18	Maas2023 (39%)	8–9	51	2.3
Internal Policies	24	NIST2024 (23%)	5–7	108	4.9
Implement	***47***			***1,577***	***71.5***
Technical Artifacts	44	Ma2025 (13%)	1–2	575	26.1
Practices & Processes	37	Crichton2025 (20%)	1–2	526	23.8
Monitoring	34	Crichton2025 (12%)	5–7	108	4.9
Evaluation & Audit	39	NIST2024 (19%)	3–3	368	16.7
Other	*32*			*171*	*7.8*
Total	47			2,206	100.0

Note. N = 2,206 mitigation records, from a total of 47 source documents. "Documents" is the number of source documents contributing at least one record to that category. For this reason, rows in the "Documents" column sum to greater than 47. Rank across deletions is the range of ranks the category took when each document in turn was removed. The largest single source and its share of the category's records indicate single-document dependence.

We conducted a robustness analysis to investigate whether the patterns of mitigations across instruments were being driven by specific documents, as with our analysis for mitigation domains. Figure 5 and Table 4 show that most mitigation instruments were stable under the robustness checks. *Technical Artifacts* and *Practices & Processes* alternated between first and second rank across single-document deletions. *External Standards* was the least robust mitigation instrument (2.3%, 95% CI [1.0, 4.4]), with 39% of its records from Maas (2023).

Figure 5. Distribution of mitigation records across mitigation instruments



Note. Intervals are 95% cluster bootstrap intervals over source documents (20,000 resamples). A hollow marker indicates that more than 20% of the category's records came from a single document, so its position in the ranking is less stable. Shares are of all eligible mitigation records, with the remainder consisting of "Other" or cross-cutting codes. N = 2,206 eligible mitigation records from 47 documents.

Distribution of mitigations by AI actor, AI lifecycle stage, AI risk addressed, and AI risk management function

In addition to the novel AI Risk Mitigation Taxonomy, we also classified mitigations against four other taxonomies to improve the usefulness of the database:

1. **AI actor** (e.g., *Developer*, *Deployer*) that implements the mitigation (adapted from Australian Government, 2024; NIST, 2023)
2. **AI lifecycle stage** (e.g., *Plan & Design*, *Build & Use Model*) where the mitigation is implemented (adapted from NIST, 2023; OECD, 2022)
3. **AI risk** (e.g., *Discrimination & Toxicity*, *Privacy & Security*) that the mitigation is intended to address, coded to both the domain and subdomain levels (Slattery et al., 2026)
4. **AI risk management function** (e.g., *Govern*, *Map*, *Measure*, *Manage*) that the mitigation is aligned with, based on the NIST AI RMF functions (NIST, 2023; OECD, 2022)

AI actor: Who implements the mitigation?

We classified mitigations against the actor in the AI ecosystem that was named as responsible for implementing or enacting the mitigation (adapted from Australian Government, 2024; NIST, 2023). Names, definitions, and distributions of mitigation records for the AI actor taxonomy are presented in Table 5. Reported mitigations were concentrated among three actor categories: *Developers* (41%), *Deployers* (33%), and *Governance Actors* (18%).

Table 5. AI actor categories in the AI Risk Mitigation Repository

Category	Definition	n	%	Documents
Developer	Entity that creates general-purpose foundation models or specialized AI systems.	894	40.5	41
Deployer	Entity that implements AI systems in products, services, or platforms used within an organization (internal deployment) or within products, services, or platforms delivered to customers or the public (external deployment).	723	32.8	41
Infrastructure Provider	Entity that provides compute, cloud infrastructure, and/or data to train and run AI.	29	1.3	12
Governance Actor	Entity that creates or enforces laws, regulations, standards or guidelines for AI development, deployment and use.	403	18.3	30
User	Entity that uses or relies on AI systems without significant modification.	23	1.0	6
Affected Stakeholder	Entity indirectly affected by AI decisions or outputs.	12	0.5	6
Other	*Mitigation describes an entity that does not fit existing codes, describes multiple entities, or cannot be classified.*	*122*	*5.5*	*25*
Total		2,206	100.0	47

Note. N = 2,206 mitigation records, from a total of 47 source documents. "Documents" is the number of source documents contributing at least one record to that category. For this reason, rows in the "Documents" column sum to greater than 47. AI actor categories adapted from Australian Government (2024) and NIST (2023).

AI lifecycle stage: When is the mitigation implemented?

We classified mitigations against the stage of the AI lifecycle where the mitigation is implemented (adapted from NIST, 2023; OECD, 2022). Details of how mitigation records are distributed across the AI lifecycle taxonomy are presented in Table 6. The largest AI lifecycle category was *Outside Lifecycle* (26%) – mitigations not inherently tied to the AI system lifecycle at all – followed by *Operate & Monitor* (19%).

Table 6. AI lifecycle stage categories in the AI Risk Mitigation Repository

Category	Definition	n	%	Documents
Plan & Design	Activities that scope the AI system's purpose, context, requirements, and preliminary risk assessment; may include prototyping.	237	10.7	27
Collect & Process Data	Acquisition, preprocessing, and preparation of the data that will be used to train the model. This takes place before the model is built and trained.	130	5.9	30
Build & Use Model	Model development, training, and initial implementation phases of AI system creation. At this stage, the model will only be used internally by developers.	258	11.7	37
Verify & Validate	Testing, evaluation and conformity checks to ensure the model meets functional, safety and fundamental requirements before release.	225	10.2	36
Deploy	Placing the AI system on the market, integrating it into operational environments, and making it available to users. This could still be a pilot phase and include checking compatibility with legacy systems and evaluation of user experience.	150	6.8	30
Operate & Monitor	Ongoing provision of the system to users after deployment, with monitoring of performance and user-experience. May include incident logging, drift detection, post-market surveillance and corrective actions throughout the system's lifetime.	426	19.3	44
Outside Lifecycle	Mitigations that are not inherently tied to the AI system lifecycle. They operate at the organizational or ecosystem level independently of any specific lifecycle stage.	576	26.1	36
Other	*Mitigation describes a lifecycle stage that does not fit into existing codes, describes multiple stages, or cannot be classified.*	*204*	*9.2*	*39*
Total		2,206	100.0	47

Note. N = 2,206 mitigation records, from a total of 47 source documents. "Documents" is the number of source documents contributing at least one record to that category. For this reason, rows in the "Documents" column sum to greater than 47. AI lifecycle stage categories adapted from OECD (2022) and NIST (2023).

AI risk: What risks are addressed by the mitigation?

We classified mitigations against the AI risk domain and/or subdomain that they addressed using the MIT AI Risk Repository domain taxonomy (Slattery et al., 2026). Names and distributions are presented in Table 7. The most common AI risk domain was *AI System Safety, Failures, & Limitations* (33%), followed by *Privacy & Security* (23%). The most common AI risk subdomain that mitigations were described as addressing was *Governance Failure* (18% of all mitigations), followed by *AI Possessing Dangerous Capabilities* (14%).

Table 7. AI risk categories in the AI Risk Mitigation Repository

Category	n	%	Documents
Discrimination & Toxicity	***133***	***6.0***	***29***
Unfair Discrimination & Misrepresentation	57	2.6	21
Exposure to Toxic Content	49	2.2	15
Unequal Performance across Groups	6	0.3	2
Discrimination & Toxicity (Other)	*21*	*1.0*	*14*
Privacy & Security	***507***	***23.0***	***41***
Compromise of Privacy by Obtaining, Leaking, or Correctly Inferring Sensitive Information	105	4.8	25
AI System Security Vulnerabilities & Attacks	316	14.3	37
Privacy & Security (Other)	*86*	*3.9*	*16*
Misinformation	***101***	***4.6***	***22***
False or Misleading Information	83	3.8	19
Pollution of Information Ecosystem & Loss of Consensus Reality	8	0.4	4
Misinformation (Other)	*10*	*0.5*	*7*
Malicious Actors & Misuse	***171***	***7.8***	***35***
Disinformation, Surveillance, & Influence at Scale	8	0.4	6
Cyberattacks, Weapons Development or Use, & Mass Harm	33	1.5	12
Fraud, Scams, & Targeted Manipulation	6	0.3	6
Malicious Actors & Misuse (Other)	*124*	*5.6*	*30*
Human-Computer Interaction	***62***	***2.8***	***19***
Overreliance & Unsafe Use	28	1.3	13
Loss of Human Agency & Autonomy	28	1.3	14
Human-Computer Interaction (Other)	*6*	*0.3*	*5*
Socioeconomic & Environmental Harm	***466***	***21.1***	***37***
Power Centralization & Unfair Distribution of Benefits	7	0.3	6
Increased Inequality & Decline in Employment Quality	7	0.3	5
Economic & Cultural Devaluation of Human Effort	12	0.5	5
Competitive Dynamics	8	0.4	5
Governance Failure	403	18.3	36
Environmental Harm	12	0.5	7
Socioeconomic & Environmental Harm (Other)	*17*	*0.8*	*11*
AI System Safety, Failures, & Limitations	***738***	***33.5***	***42***
AI Pursuing Its Own Goals in Conflict with Human Goals or Values	77	3.5	23
AI Possessing Dangerous Capabilities	300	13.6	31
Lack of Capability or Robustness	128	5.8	21
Lack of Transparency or Interpretability	108	4.9	27
AI Welfare & Rights	0	0.0	0
Multi-agent Risks	17	0.8	7
AI System Safety, Failures, & Limitations (Other)	*108*	*4.9*	*24*
Other	*28*	*1.3*	*13*
Total	2,206	100.0	47

Note. N = 2,206 mitigation records, from a total of 47 source documents. "Documents" is the number of source documents contributing at least one record to that category. For this reason, rows in the "Documents" column sum to greater than 47. AI risk categories from Slattery et al. (2026).

NIST AI risk management function: Which risk management function is aligned with the mitigation?

We classified mitigations according to the function they most closely aligned with in the NIST AI Risk Management Framework (NIST, 2023; OECD, 2022), which aligns with many other enterprise risk management frameworks in structuring mitigation practices into *Govern*, *Map*, *Measure*, and *Manage* functions. Table 8 presents the names and definitions of the categories, and the distribution of mitigations against them. The most reported mitigations were aligned with *Manage* (37%), followed by *Govern* (31%).

Table 8. NIST AI Risk Management Framework function categories in the AI Risk Mitigation Repository

Category	Definition	n	%	Documents
Govern	A culture of risk management is cultivated and present.	694	31.5	40
Map	Context is recognized and risks related to context are identified.	170	7.7	23
Measure	Identified risks are assessed, analyzed, or tracked.	474	21.5	40
Manage	Risks are prioritized and acted upon based on a projected impact.	827	37.5	43
Other	*Mitigations that do not fit the functions above or cannot be classified.*	*41*	*1.9*	*14*
Total		2,206	100.0	47

Note. N = 2,206 mitigation records, from a total of 47 source documents. "Documents" is the number of source documents contributing at least one record to that category. For this reason, rows in the "Documents" column sum to greater than 47. NIST AI RMF function categories from NIST (2023).

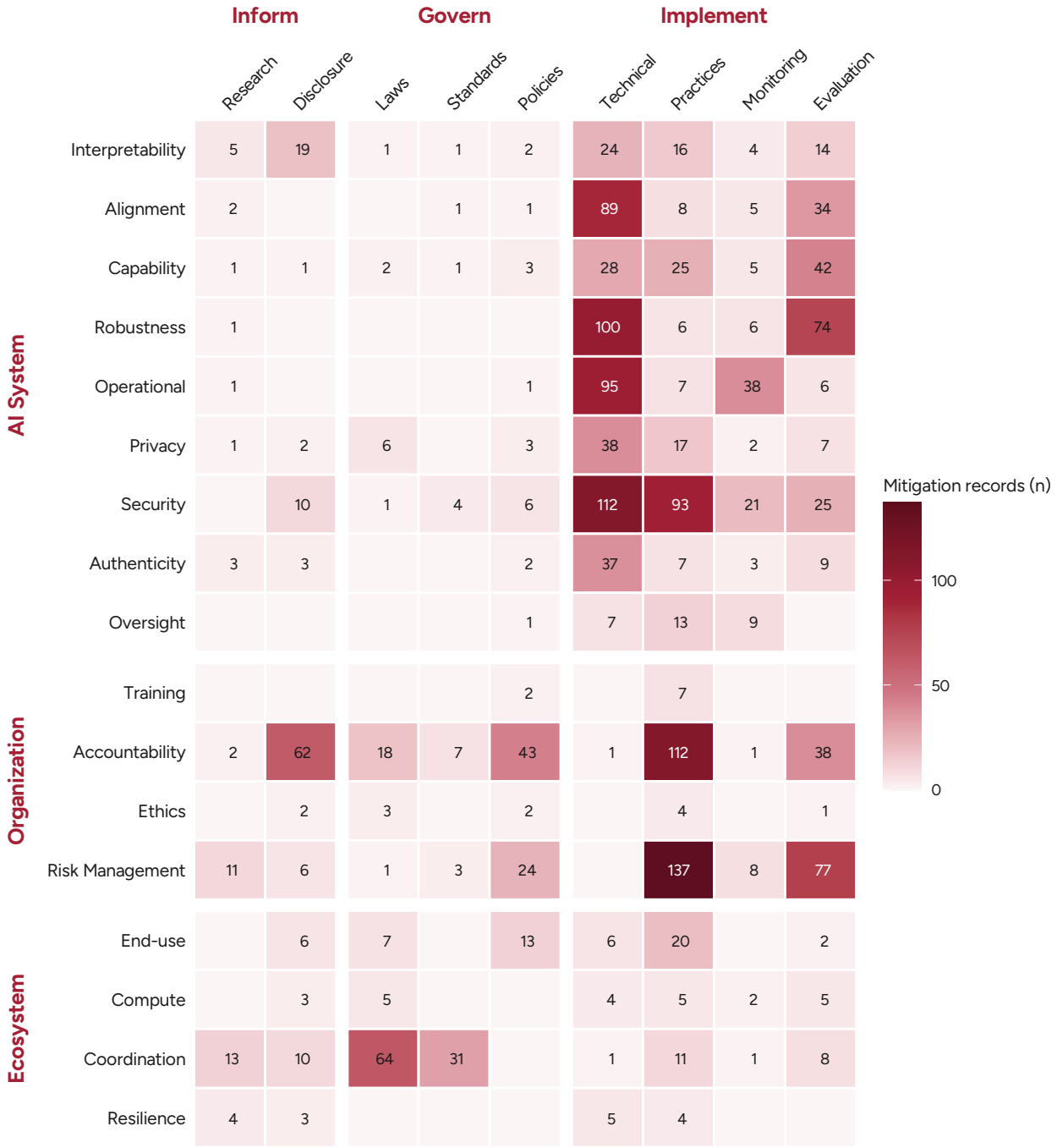
AI risk mitigations across multiple dimensions: *Who* does *what* and *when* to address *which* risks?

Bringing together all taxonomies, we examined how mitigations were commonly represented through combinations of mitigation domains, instruments, actors, lifecycle stages, and risks addressed. Figure 6 shows a heatmap of mitigation records from the repository distributed across mitigation domains and instruments. There were clear differences by locus: in the *AI System* locus, *Technical* and *Evaluation* instruments dominated, while in the *Ecosystem* locus (and specifically its *Coordination* domain) *Law & Regulation* and *External Standards* dominated. Even so, the *Ecosystem* locus had the most even spread across the three functions (*Inform* 17%, *Govern* 52%, *Implement* 32%). The heatmap also shows where domain–instrument combinations are sparse, with 42 of 153 containing no records. *Inform* and *Govern* instruments were rare for AI system domains, but not absent (e.g., "model cards" and "documentation" as *Information Sharing & Disclosure* instruments for *Interpretability & Transparency*; "data protection laws" as *Laws* for *Privacy & Data Protection*). *Technical* instruments dropped sharply beyond the AI system but still appeared (e.g., "pathogen detection tools" for *Societal Resilience & Recovery*).

Which categories of mitigations are described as addressing which risks from AI?

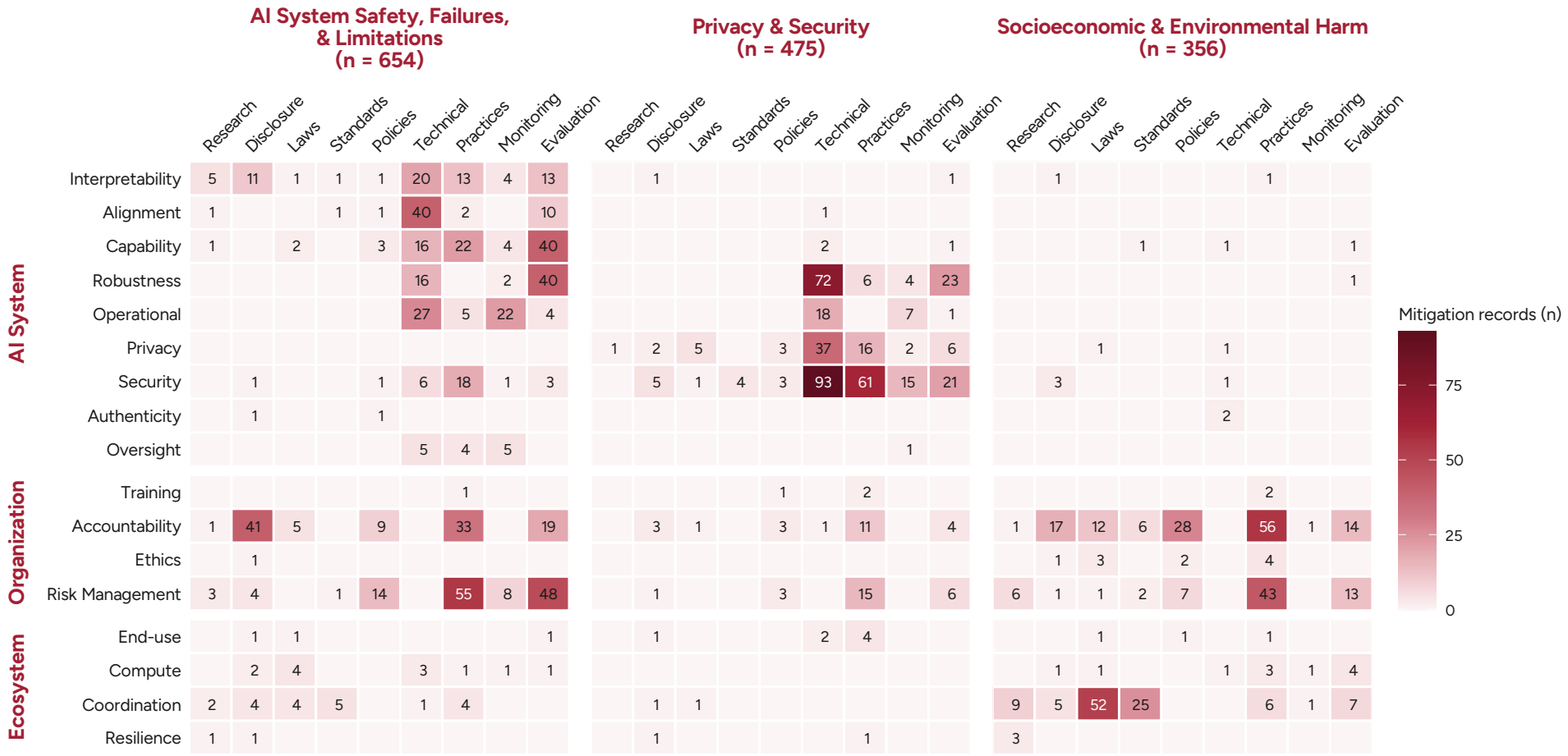
Figure 7 shows the same heatmap of mitigation domains and instruments, broken down by the three top risk domains addressed: *AI System Safety, Failures, & Limitations*; *Privacy & Security*; and *Socioeconomic & Environmental Harm*. The findings show clear interactions with risk addressed: mitigations in response to *Socioeconomic & Environmental Harm* concentrated in the *Organization* and *Ecosystem* loci, while conversely, mitigations addressing *Privacy & Security* concentrated in the *AI System* locus and in *Technical* instruments. Mitigations addressing *AI System Safety, Failures, & Limitations* showed a mix of implementing instruments in the AI system as well as organization-level *Accountability* and *Risk Management* strategies.

Figure 6. Distribution of mitigation records by mitigation domain and instrument



Note. Color of cell indicates relative frequency. Mitigation domains and instruments are shown by their short labels. N = 1,916 records from 47 documents. Excludes risk mitigation records coded to "Other" categories.

Figure 7. Distribution of mitigation records by mitigation domain and instrument for the three largest domains of risk addressed



Note. Color of cell indicates relative frequency. Mitigation domains and instruments are shown by their short labels. N = 1,485 records from 46 documents. Excludes risk mitigation records coded to "Other" categories. The risk domains shown are the three largest categories in the mitigation repository.

What actors are described as responsible for developing, implementing, and/or governing each category of mitigations?

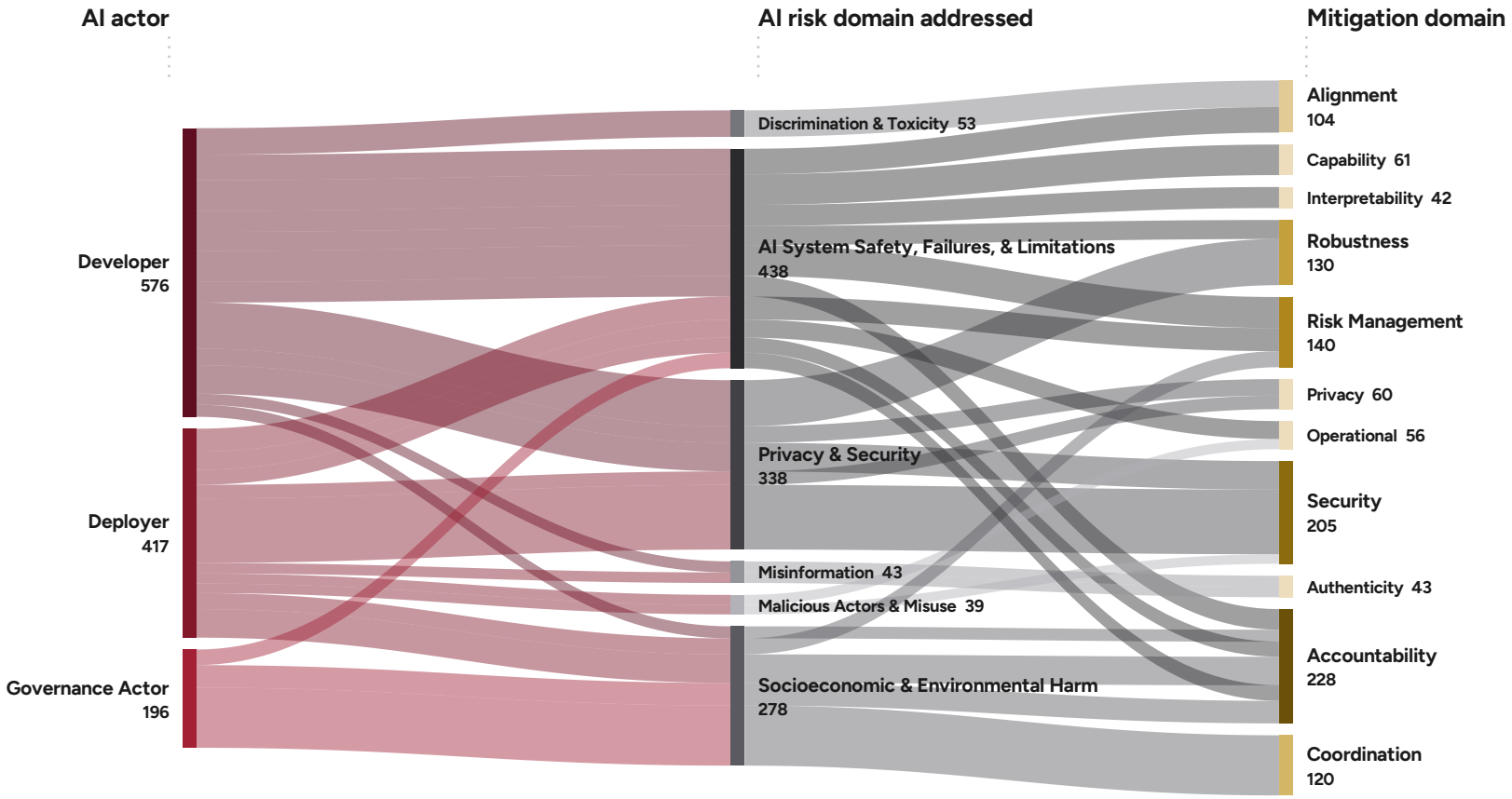
The most commonly appearing links between AI actors, risks addressed, and categories of mitigations in the literature are presented in the Sankey diagram of Figure 8, where band width is proportional to the number of mitigation records. This shows the 25 most common AI actor–risk–mitigation domain combinations, representing 60% of records specified on all three taxonomies. Three actors dominated: *Developers* were associated chiefly with addressing risks from *Privacy & Security* and *AI System Safety, Failures, & Limitations*, largely through a range of AI system mitigation domains (*Alignment & Value Learning*, *Robustness & Adversarial Resistance, Capability Control & Containment*). *Deployers* were associated mainly with *Privacy & Security*; *AI System Safety, Failures, & Limitations;* and *Socioeconomic & Environmental Harm*. *Governance Actors* predominantly addressed *Socioeconomic & Environmental Harm* through *Coordination* and *Accountability* mitigations.

Figure 9 shows the distribution of mitigation domains across lifecycle stages, where column width is proportional to the number of mitigation records for a given stage, and segment height is the mitigation domain's share of that stage. Within the lifecycle, AI system domains (*Robustness*, *Security*, *Operational Safety & Control*) dominated, except for early-stage *Plan & Design*, where *Risk Management* mitigations took more than half the share. Mitigations outside the lifecycle were common, at over a quarter of specified records. These were dominated by *Coordination* (31% of mitigations outside the lifecycle) and *Accountability* (27%)[4].

Among instruments (Figure 10), implementing instruments (*Technical, Practices, Evaluation, Monitoring*) dominated every lifecycle stage. However, this was no longer the case outside the lifecycle, where records were more evenly dispersed across *Govern* (44%, including *Laws*) and *Implement* (41%) instruments.

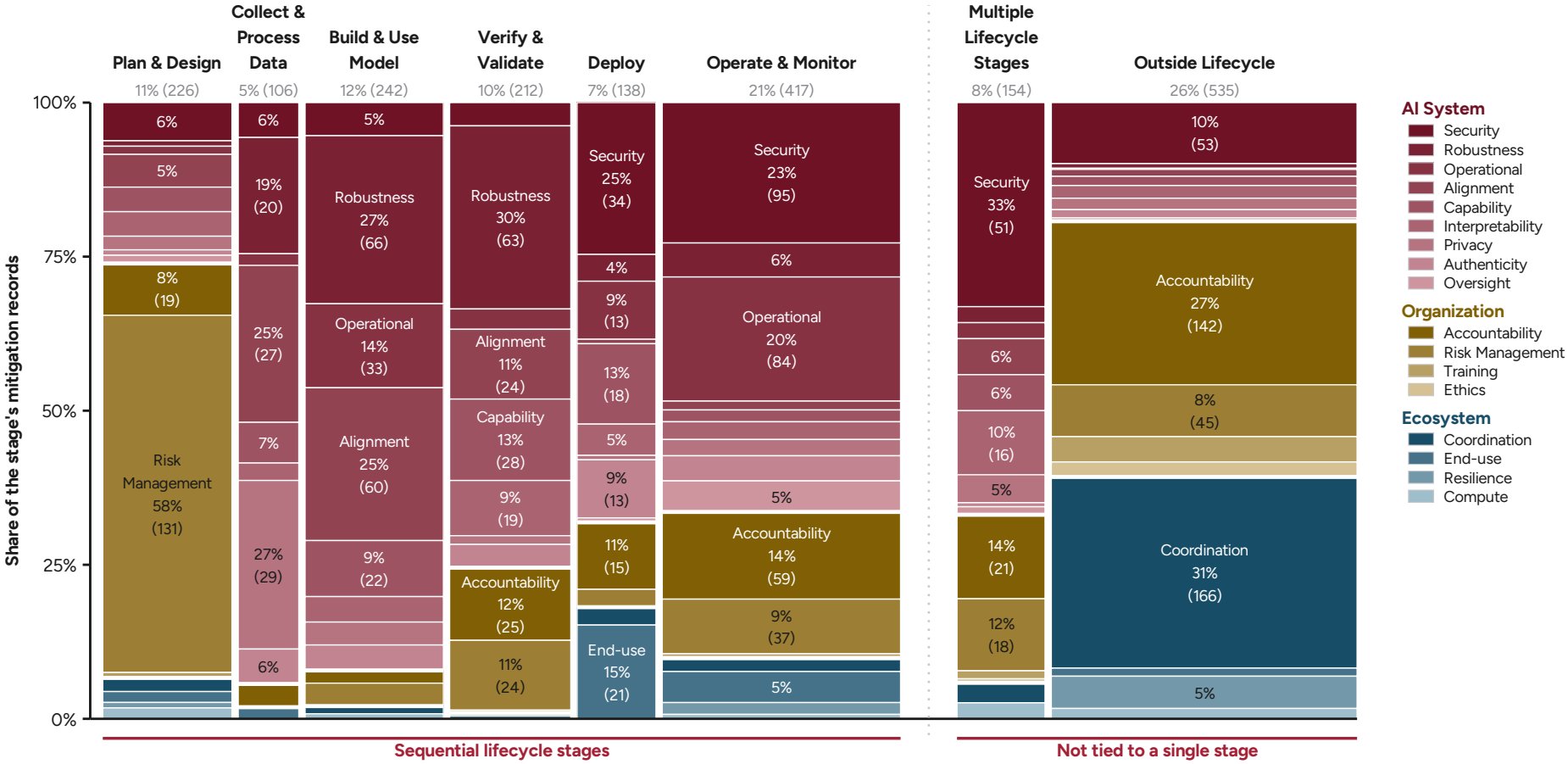
[4] Given the low robustness for *Coordination,* we repeated this analysis using distinct contributing documents as the unit of analysis. The share of mitigations outside the lifecycle dropped to 15% for both *Coordination* and *Accountability*, although the pattern of findings described above remained largely the same.

Figure 8. Most frequent AI actor–risk domain–mitigation domain combinations in the AI Risk Mitigation Repository



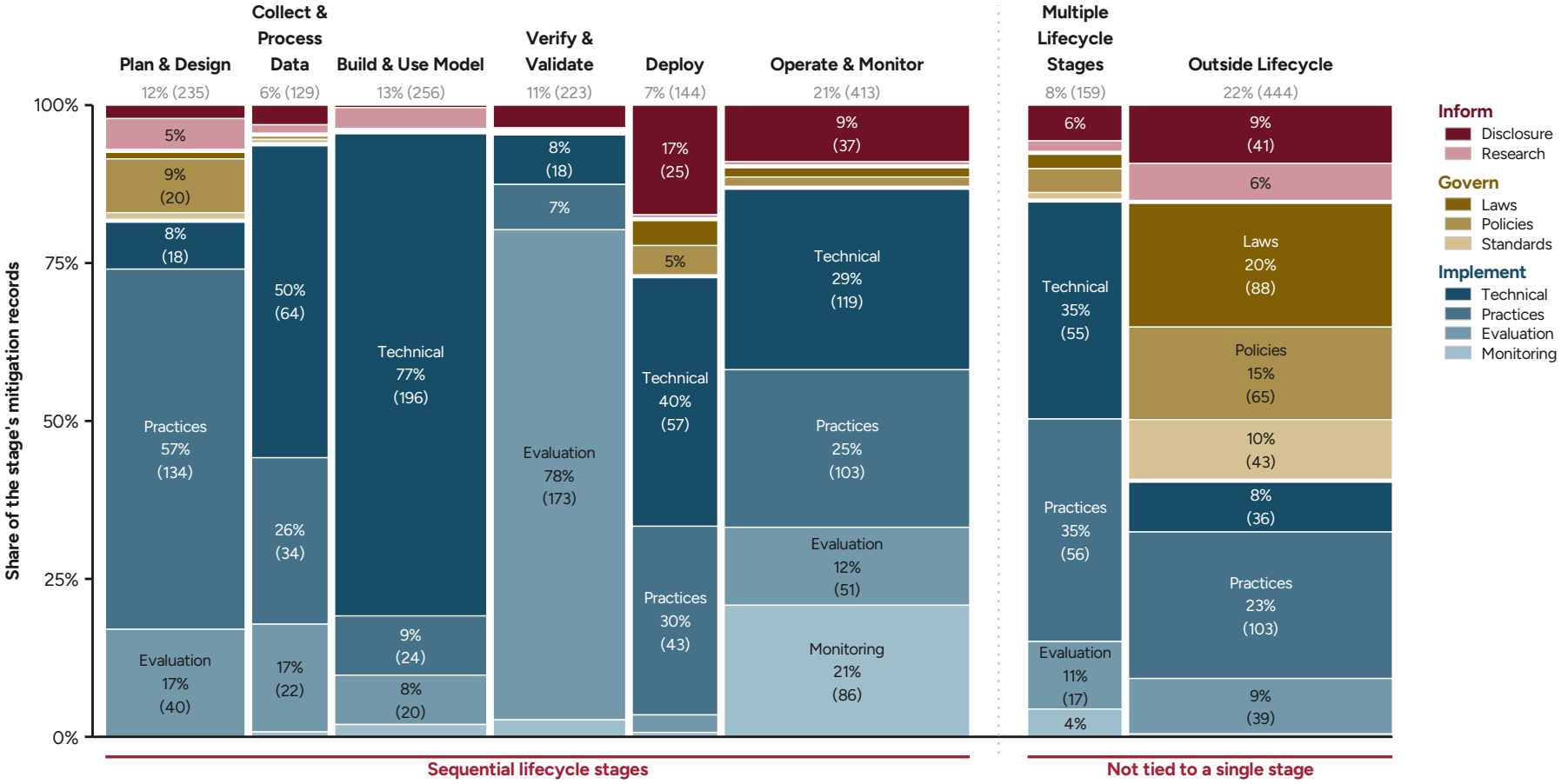
Note. The 25 most common combinations, representing 60% of records where all three taxonomies were coded, and 54% of all records. Mitigation domains are shown by their short labels.

Figure 9. Distribution of mitigation records by AI lifecycle stage and mitigation domain



Note. Column width reflects the number of mitigation records for that stage; segment height is the domain's share of that stage, so each column sums to 100% and cell area reflects cell count. Larger cells show the category's short label, its within-stage percentage and its record count; smaller cells show the percentage and count, or the percentage alone, where they fit. N = 2,030 records specified on both dimensions (92% of 2,206 eligible records from 47 documents).

Figure 10. Distribution of mitigation records by AI lifecycle stage and mitigation instrument



Note. Column width reflects the number of mitigation records for that stage; segment height is the instrument's share of that stage, so each column sums to 100% and cell area reflects cell count. Larger cells show the category's short label, its within-stage percentage and its record count; smaller cells show the percentage and count, or the percentage alone, where they fit. N = 2,003 records specified on both dimensions (91% of 2,206 eligible records from 47 documents).

Discussion

Overview of findings

Our systematic review of AI risk mitigations identified 47 relevant documents and established a repository of 2,206 AI risk mitigation records. Each mitigation was classified against five taxonomies: a new two-axis taxonomy of AI risk mitigation domains and instruments, as well as AI actor, AI lifecycle stage, AI risk addressed, and AI risk management function taxonomies. To our knowledge, this is the first project to present the landscape of AI risk mitigations on this scale and across five taxonomies, and to share the data for further use. Connecting these dimensions gives a fuller picture of what mitigations are proposed and supports implementation by helping researchers and practitioners understand mitigations based on who should act, when, against which risks, and through what method.

In this section we (1) describe how mitigations are distributed across the novel AI Risk Mitigation Taxonomy and how this may reflect the evolving field of AI risk mitigations; (2) share what we learned from developing a novel taxonomy and classifying mitigations; (3) propose a reporting standard for mitigations to improve consistency and usefulness of AI risk mitigation reporting; (4) describe how the taxonomies and repository can be used; and (5) describe limitations and future directions for this work.

Insights from the literature: What clusters and gaps across taxonomies reveal about mitigations

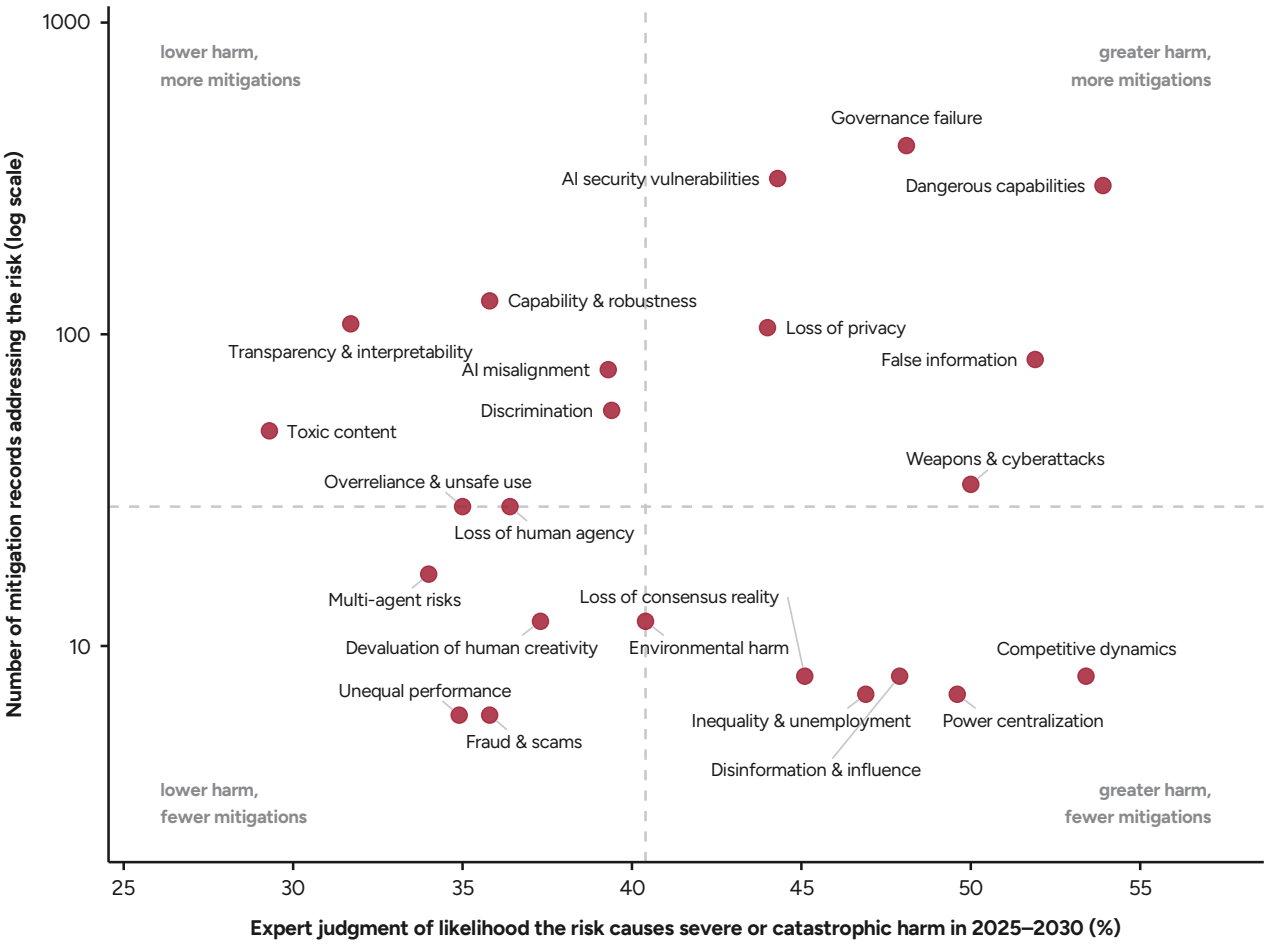
Our analysis of the AI risk mitigation literature revealed that the stereotypical AI risk mitigation is a *technical artifact* implemented *at the AI system level* by an *AI developer* during the *AI system lifecycle*, aligned with *measuring* or *managing* an AI risk. Almost every included document contained *AI System* mitigations (98% of documents), compared with *Human & Organization* (77%) and *Society & Ecosystem* (72%).

Some domains of mitigations were much less common. Five domains each accounted for less than 2% of mitigation records: *Integrity & Ethics*; *Training & Capacity-building*; *Compute, Hardware, & Resource Stewardship*; *Human Oversight & Agentic Control*; and *Societal Resilience & Recovery*. Most of these domains were introduced through expert consultation, in which experts identified gaps in coverage of human oversight and agentic control, end-use governance, societal resilience, and ex post liability.

In an exploratory analysis, we compared the most commonly addressed AI risks in our review with expert judgments of priority AI risks (Saeri et al., 2026). Figure 11 plots the number of mitigation records addressing each risk against its expert-rated probability of severe or catastrophic harm. Many risks fell where expected, with greater-harm risks such as *Governance Failure* and *Dangerous Capabilities* well represented by mitigation records, and lower-harm risks such as *Unequal Performance* and *Fraud & Scams* less so. The exceptions were *Competitive Dynamics*

and *Power Centralization*, among the five most severe risks in Saeri et al. (2026), which were addressed by only a few mitigation records (15, or under 1% of the repository)[5].

Figure 11. Expert-rated probability of severe or catastrophic harm against mitigation record count, by AI risk subdomain



Note. Each point is one AI risk subdomain (presented using short labels) from the MIT AI Risk Repository domain taxonomy. The horizontal axis is the expert judgment of the likelihood the risk causes severe or catastrophic harm in 2025–2030 under a business-as-usual scenario, from a Delphi study of 272 experts (Saeri et al., 2026). The vertical axis is the number of mitigation records in the repository addressing the risk (log scale). Dashed lines mark the medians of the two axes. N = 1,806 eligible mitigation records with risk addressed specified.

[5] This pattern is robust to different ways of measuring attention to mitigations and risk prioritization. Attention to mitigations can also be measured through (1) the number of mitigation records that mention this risk, shown in Figure 11, or (2) the number of distinct documents that mention this risk. Risk prioritization can be measured through (1) the likelihood that the risk causes severe or catastrophic harm, shown in Figure 11, or (2) ratings of severity associated with the risk. *Competitive Dynamics* and *Power Centralization* lie above the median for risk prioritization and below the median for attention to mitigations on all four combinations of these measures.

Classifying AI risk mitigations: Learnings and challenges

In this section, we discuss what we learned about developing a novel taxonomy of AI risk mitigations, and the challenges of classifying more than 2,000 diverse mitigation records against existing and novel taxonomies.

Separating mitigation domains and instruments

As we iterated on the development of a novel taxonomy of AI risk mitigations, we decided to separate mitigation *domains* (what a mitigation aims to improve, e.g., interpretability, oversight) from mitigation *instruments* (the levers through which a mitigation acts, e.g., law, technical artifact). Existing frameworks have typically organized mitigations along a single dimension, such as the AI RMF function (Barrett et al., 2025), lifecycle stage (Joshi et al., 2025), or along two or three dimensions (pathways to influence, policies, and AI actors; Maas, 2023; locus, lifecycle stage, and governance components; Mäntymäki et al., 2022; risk area, lifecycle stage, and technical vs. process treatments; OECD, 2023).

We separated the domain and instrument axes because the mitigation literature itself often described mitigations using similar terms like "interpretability" and "whistleblower protections", but in fact referred to different levers or domains of change. The *instrument* axis was also intended to connect mitigations proposed for AI to generic levers for change that apply beyond AI, inspired by other intervention taxonomies for societal challenges. Hood's (1983) classification of government resources, the policy categories of the Behaviour Change Wheel (Michie et al., 2011), and the Haddon matrix (Haddon, 1972) all pair a set of generic levers with a second orthogonal axis.

The value of this approach is that recognizable interventions stay intact in one axis while varying on the other (for other similar approaches, see also Howlett, 2009; Strom et al., 2020). For example, "watermarking" includes several mitigations that can mostly be classified under the domain *Content Authenticity & Provenance*, but span several instruments. Most "watermarking" mitigations are described as using the instrument *Technical Artifacts* (e.g., "embed watermarks […] into the model's outputs"), but also include *Research & Evidence* ("invest in researching how AI-generated content may be watermarked"), *Internal Policies* ("establish policies for measuring the effectiveness of employed content provenance methodologies"), and *Evaluation & Audit* ("measure reliability of content authentication methods"). In contrast, "red teaming" can almost always be classified under the instrument *Evaluation & Audit*, but may span several domains, including *Robustness & Adversarial Resistance, Security & Access Control,* and *Privacy & Data Protection*.

Some clusters of mitigations included multiple domains and instruments. For example, proposals to mitigate AI risks by stopping, pausing, or "pacing" AI development had multiple embedded components. Specifically, pausing a training run or setting intolerable risk thresholds were most often classified under the domain *Capability Control & Containment*, but a coordinated pause among developers would be classified under the domain *Ecosystem Coordination*. Instruments for an AI pause could include *Information Sharing & Disclosure* (as in an open letter calling for a pause), *Law & Regulation* (if a pause was mandated), or *Internal Policies* (if an AI developer unilaterally adopted a pause).

Classifying diverse mitigation records

Classifying records against the taxonomies raised three challenges: the boundary of what counts as a mitigation, the heterogeneity of the source records, and the sensitivity of classification to wording.

We intentionally chose a broad definition of AI risk mitigations ("an action that reduces the likelihood or impact of a risk"; after Actuarial Standards Board, 2014; Aven et al., 2018) to admit a wide range of actions and documents. However, this led to challenges in defining the boundary of a mitigation in practice, since documents varied widely in how they determined whether a concept was a mitigation. Our expert panel also disagreed about the relationship between mitigations and risk management measures. Some risk management functions, such as *Map* and *Measure*, which identify and analyze risks rather than act on them, may not be conventionally considered mitigations. We retained them in the repository because these functions cannot always be separated in practice. For example, a red-teaming evaluation that identifies a vulnerability may also test whether an existing safeguard holds.

Mitigation records varied greatly in specificity and form, from "theoretical foundations" and "PhotoDNA" to "service providers should publicize capabilities, limitations, target users, and use cases of AI products and services". Some were described by the risk they were intended to address, rather than the action taken to address that risk (e.g., "improve model security" addresses security risks by definition). We addressed this by excluding records that were too vaguely worded to produce a meaningful classification, or provided only structural information (e.g., a "security-enhancing mitigations" parent category for four specific mitigations, each intended to reduce model security vulnerabilities). We also relied on contextual information from the document. Even so, records varied in how consistently they were defined, blurring the line between core interventions and other factors such as tools and techniques (e.g., "Delphi method") and enabling conditions (e.g., "safety culture"). Core interventions act on the risk pathway directly; techniques and tools are means of delivery that act only through an intervention; and enabling conditions are states rather than actions. Future work could establish rules to separate the three, drawing on precedents for rigorous reporting on interventions from other areas (Hoffmann et al., 2014).

We aimed to preserve the fidelity of mitigations as presented by authors, but this revealed how sensitive categorization could be to subtle differences in wording. "Invest in researching how AI-generated content may be watermarked", "incentivize AI research effort", and "enabling the ecosystem with research funding" could each be coded differently depending on whether the active ingredient was economic incentives, the research itself, or even the objective of the research. There was no standardization in naming or description of mitigations: only 3.4% of mitigation records in the repository shared an exact name with another (e.g., "adversarial training").

A proposed mitigation reporting standard

Our challenges in classifying mitigations represent a broader problem in the literature: AI risk mitigations are not reported consistently. Across documents, mitigations varied widely in granularity, type of action, mechanism, and intended outcome, and did not have standard names. Similar issues have previously been identified in health intervention research, where incomplete

descriptions of health interventions made trial results difficult to replicate or compare, a major source of research waste (Chalmers & Glasziou, 2009; Glasziou et al., 2014; Michie et al., 2009). That research area has since adopted minimum reporting standards, including CONSORT for trials (Consolidated Standards of Reporting Trials; Hopewell et al., 2025) and TIDieR for interventions (Template for Intervention Description and Replication; Hoffmann et al., 2014). TIDieR is a 12-item checklist for how interventions should be described, including name, rationale, materials, procedures, and mode of delivery. These standards aim to improve the clarity, completeness, and transparency of reporting, which is a precondition for cumulative evidence on what works.

We propose an equivalent reporting standard for AI risk mitigations: PRIMER (Preferred Reporting Items for Mitigations of Emerging Risks; Figure 12). PRIMER includes eight items based on the taxonomies in the mitigation repository: mitigation name, mitigation role, risk addressed, objective, mechanism, actor, timing, and linked mitigations. This covers four of the seven dimensions that have been recommended for specifying implementation strategies: actor, action, action target, and temporality (Proctor et al., 2013). Figure 13 gives a worked example of PRIMER using a mitigation record from the repository ("Define risk thresholds"). PRIMER is designed for public reporting, but complements existing structured mitigation and control catalogs, such as Google's (2024) Secure AI Framework Risk Map, which describes controls by title, description, category, personas, components, and risks; and Credo AI's Unified Control Framework (Eisenberg et al., 2025; see also Cloud Security Alliance, 2026; Derczynski et al., 2023).

When using PRIMER to report on mitigations, we recommend two additional checks: (1) *specificity*, that the mitigation cannot be decomposed into more than one core intervention, and (2) *observability*, that the mitigation is an action that can be observed and measured. AI developers and deployers using PRIMER for transparency could also report the implementation status of the mitigations they use to address AI risks (e.g., whether the mitigation is planned, piloted, or implemented). We offer PRIMER as a starting point. Future work could refine the standard through expert feedback and consensus (as CONSORT and TIDieR were developed; Moher et al., 2010), and test whether it improves consistency and comparability of mitigation classification and research, including agreement between independent coders.

Figure 12. PRIMER: Preferred Reporting Items for Mitigations of Emerging Risks

PRIMER

Preferred Reporting Items for Mitigations of Emerging Risks

MIT AI Risk Initiative

GROUP	#	ITEM	QUESTION IT ANSWERS	RESPONSE FIELDS
IDENTITY	1	**Mitigation name**	What is the mitigation called?	
	2	**Mitigation role**	Is this a core intervention, a supporting technique, or an enabling condition?	Core intervention Supporting technique Enabling condition
PURPOSE	3	**Risk**	What risk does this mitigation address?	[Risk domains / subdomains]
	4	**Objective**	What does this mitigation improve if it succeeds?	[Mitigation domains]
OPERATION	5	**Mechanism**	Through what lever does this mitigation act?	[Mitigation instruments]
	6	**Actor**	Who carries out this mitigation?	[AI actor]
	7	**Timing**	When does this mitigation take place?	[AI lifecycle stage]
	8	**Linked mitigations**	What mitigations come before this, and what comes after it, for it to succeed?	

TWO CHECKS BEFORE A MITIGATION IS COMPLETE

Specificity: The mitigation cannot be decomposed into more than one core intervention.

Observability: The mitigation is described as a clear action that can be observed and measured. Excludes statements of intent, aspirations, or general principles.

Figure 13. PRIMER worked example

PRIMER

Worked example: "Define risk thresholds" from *Emerging processes for Frontier AI Safety* (UK Department for Science, Innovation and Technology, 2023)

MIT AI Risk Initiative

#	ITEM	RESPONSE FIELDS
1	**Mitigation name**	***Define risk thresholds*** *Why: The name used in the source document.*
2	**Mitigation role**	**Supporting technique** *Why: Thresholds reduce no risk themselves; they act only through the interventions they trigger (e.g., pausing, safeguards).*
3	**Risk**	***Domain: AI System Safety, Failures, & Limitations • Subdomain: AI Possessing Dangerous Capabilities*** *Why: Thresholds target dangerous capabilities, e.g., cyberattacks.*
4	**Objective**	***Locus: Human & Organization • Domain: Risk Management & Assurance*** *Why: Sets the level of AI risk the organization will accept.*
5	**Mechanism**	***Function: Govern • Instrument: Internal Policies*** *Why: An internal rule on what to do when a threshold is reached.*
6	**Actor**	***Developer*** *Why: Set by the frontier AI organization that develops the model.*
7	**Timing**	***Plan & Design*** *Why: Set before development proceeds, as criteria for continuing.*
8	**Linked mitigations**	***Before*** • Devote resources to pre-development risk assessments • Engage external stakeholders when developing risk thresholds ***After*** • At each threshold, commit to proceed only with adequate mitigations • Inform government when a threshold is met • Prepare to pause training or reduce access

TWO CHECKS BEFORE A MITIGATION IS COMPLETE

Specificity

✓ PASS
Why: One action; operationalizing and acting on them are separate linked mitigations.

Observability

✓ PASS
Why: Produces a documented set of thresholds that can be inspected.

Worked example

PRIMER items

Practical implications and future directions

The AI Risk Mitigation Repository, the novel AI Risk Mitigation Taxonomy, and the overall multi-taxonomy framework are designed for users across the AI ecosystem to address risks from AI.

AI developers and deployers can use the repository as a search space to discover mitigations appearing in the literature but not in current practice. Records can be filtered by actor, lifecycle stage, and risk, and the resulting list read against what an organization already does. The crossed domain and instrument axes also function as a coverage map: where an organization's mitigations for a given risk fall entirely within one instrument, the remaining cells show what other kinds of levers could be considered for the same problem.

Policymakers and regulators can use the instrument axis of the AI Risk Mitigation Taxonomy to explore the range of levers that have been proposed or applied to address AI risks, including laws, technical artifacts, evaluations, and information-sharing arrangements. The PRIMER reporting standard offers a ready-made template for reporting mitigations for AI risks. PRIMER could complement or be embedded within the HAIP Reporting Framework (OECD, 2025), which requires organizations to submit mitigation information in narrative form (see also the EU General-Purpose AI Code of Practice; European Commission, 2025).

For **researchers** of AI risks and mitigations, the mitigation repository is a rich resource that can be analyzed to understand the landscape of AI risk mitigations. We recommend that researchers describe mitigations using PRIMER in their research outputs. The multi-taxonomy framework offers a common language for classifying and comparing mitigations from other corpora, including laws and regulations, company disclosures, incident reports, organizational strategies, and funding proposals, to understand the mitigations being proposed or implemented. A further use is to compare supply against harm, for example, drawing on incident trackers to map where mitigations are concentrated relative to where damage from AI is occurring.

We identify three directions for future research. First, future research could better understand relationships and chains between mitigations. When we extracted and classified mitigations, we treated each record as an isolated action, though mitigations are rarely implemented or act individually. For example, setting intolerable risk thresholds requires a series of actions, such as a third party defining capability thresholds, a developer assessing models against them, and the developer preparing to halt development and deployment on breach. Reported as one mitigation, it fails the specificity check; reported as three or more linked mitigations, each is observable and separately implementable. The PRIMER reporting standard includes a "linked mitigations" item to record this relationship. Mapping these relationships would show where mitigations may enhance, compensate for, or undermine each other, and identify critical components where a single missing link could leave a risk unmitigated.

Second, the taxonomies describe what a mitigation targets and how it acts, but not whether it works, what it costs, whether it is feasible, or what unintended effects it may have. This information is critical for deciding which mitigations to prioritize. Future research should assess mitigations against these criteria, starting with those addressing the most pressing risks.

Third, the repository and PRIMER need to be tested in use. PRIMER should be piloted with organizations that author mitigation frameworks, to establish whether reporting against its items is feasible and whether it improves the comparability of new proposals. The multi-taxonomy

framework would benefit from user testing with practitioners, to establish which combinations of taxonomies can inform decisions about what to implement.

The AI Risk Mitigation Taxonomy was derived from documents published to early 2026 and from expert consultation. We expect the taxonomy to be revised as practices evolve and new mitigations are proposed, and welcome suggestions of further frameworks and mitigations to be included so that the repository can be a living resource. We also welcome work by researchers to extend the repository by classifying mitigation records against additional taxonomies.

Limitations

This work has four main limitations: inability to prioritize mitigations; representativeness of the included documents; tradeoffs in the construction and application of the taxonomies; and reliability issues in taxonomy classification.

The mitigation repository maps the mitigation landscape. However, it does not describe which mitigations are implemented, more effective, or should be prioritized. Like a map, the repository and taxonomies can therefore be read as a detailed description of what exists, but do not by themselves indicate the best route for addressing AI risks.

Our results are not representative of all mitigations in the literature and in practice. The findings were based on a systematic review of academic and gray literature, and source documents did not contribute an equal number of mitigation records. Our robustness checks found that most mitigation domains and instruments moved by no more than one place after excluding the highest-contributing document. Around half of the documents were review papers (23 of 47), raising the potential for double counting, which would inflate a mitigation category's apparent frequency. AI risk mitigations in practice include generic or existing mitigations that happen to address AI risks (e.g., existing know-your-customer controls may address AI fraud and scams, even if not designed to do so). AI developers, deployers, and users may also mitigate AI risks through actions consistent with obligations, requirements, and implementation guidance that lie outside our inclusion criteria at the document level. For example, we did not conduct an exhaustive search of system standards (e.g., ISO/IEC 42001), laws and regulations (e.g., Colorado AI Act), software toolkits (e.g., IBM AI Fairness 360), and curated tool catalogs (e.g., OECD's Catalogue of Tools & Metrics for Trustworthy AI). The repository is therefore not an exhaustive inventory of mitigations but a systematic account of how mitigations are proposed and organized in the literature.

We faced many tradeoffs in the construction and application of the taxonomies. We recognize that taxonomies should be concise, robust, comprehensive, extendible, and explanatory, as well as mutually exclusive and collectively exhaustive (Nickerson et al., 2013). But in practice no working classification can simultaneously achieve all of these (Bowker & Star, 1999), and our taxonomies are the result of several tradeoffs against these criteria. For instance, increasing the breadth of mitigations with the two-axis structure meant a loss of depth found in an earlier draft of the taxonomy. Combined with inconsistencies in how mitigations were described, this meant that

some mitigation records were assigned to *Other* categories[6], ranging from 1.7% (lifecycle stage, excluding multiple stages or outside the lifecycle) to 18% (AI risk).

Inter-rater reliability for mitigation classification varied across the five taxonomies. Although LLM-human agreement matched or exceeded human-human agreement for all taxonomies, Cohen's kappa ranged from "slight" to "moderate" (LLM-human κ = 0.21 to 0.59; human-human κ = 0.11 to 0.50; Landis & Koch, 1977). The lowest agreement was for the AI risk addressed by a mitigation, which we hypothesize reflects conceptual overlap between subdomains in the AI risk taxonomy (e.g., dangerous capabilities vs. AI misalignment) and a lack of information in mitigation records about the risks they targeted. Where mitigation names alone were uninformative, we classified records using a combination of available information, including definitions, examples, document context, and sibling mitigations. More generally, agreement was constrained by both the coding scheme and source material, including inconsistencies and incompleteness in how mitigations are reported. We made the choice to avoid fine-tuning category definitions and boundary criteria to maximize coder agreement. We believed that doing so would have overfitted categories to the records, undermined conceptual coverage, and undercut the expert input that shaped the categories.

Conclusion

There are many proposed solutions for mitigating AI risks, but they are difficult to navigate and compare. We developed a two-axis taxonomy that distinguishes what a mitigation aims to change (domain) from the lever it uses (instrument). We classified 2,206 mitigation records against five taxonomies: our novel taxonomy, who implements it (actor), when it happens (lifecycle), what risk it addresses (risk), and the risk management function it serves. Mitigations concentrated at the level of the AI system, and on the actions of AI developers and deployers, but were described inconsistently and often lacked detail. We propose PRIMER, a reporting standard to improve consistency and comparability in AI risk mitigation research and practice.

Two conditions must be met to move from describing and categorizing mitigations to confidently implementing the most effective ones. The first is consistent reporting of mitigations, which PRIMER is intended to support, so that mitigations can be compared and synthesized. The second is evidence of effectiveness, cost, and implementation feasibility, so that mitigations can be prioritized.

We publish the repository, taxonomy, and reporting standard openly to give researchers, developers, deployers, and policymakers a shared map and a common language. Mitigation researchers and authors of mitigation frameworks should report against PRIMER; researchers, developers, and deployers should compare and evaluate mitigations; and policymakers and regulators should require developers and deployers to report their mitigations consistently and show evidence that they work. The task now is to establish which mitigations work, and implement them to reduce harm from the most pressing AI risks.

[6] Mitigations were classified in *Other* categories for 1.7% of AI lifecycle stages (excluding multiple stages or outside the lifecycle), 2.7% of AI actors (excluding multiple actors), 1.9% of AI risk management functions, 6.3% of mitigation domains, 7.8% of mitigation instruments, and 18% of risks addressed. For mitigation domains and risks addressed, most of the imprecision was at more granular levels of categorization.

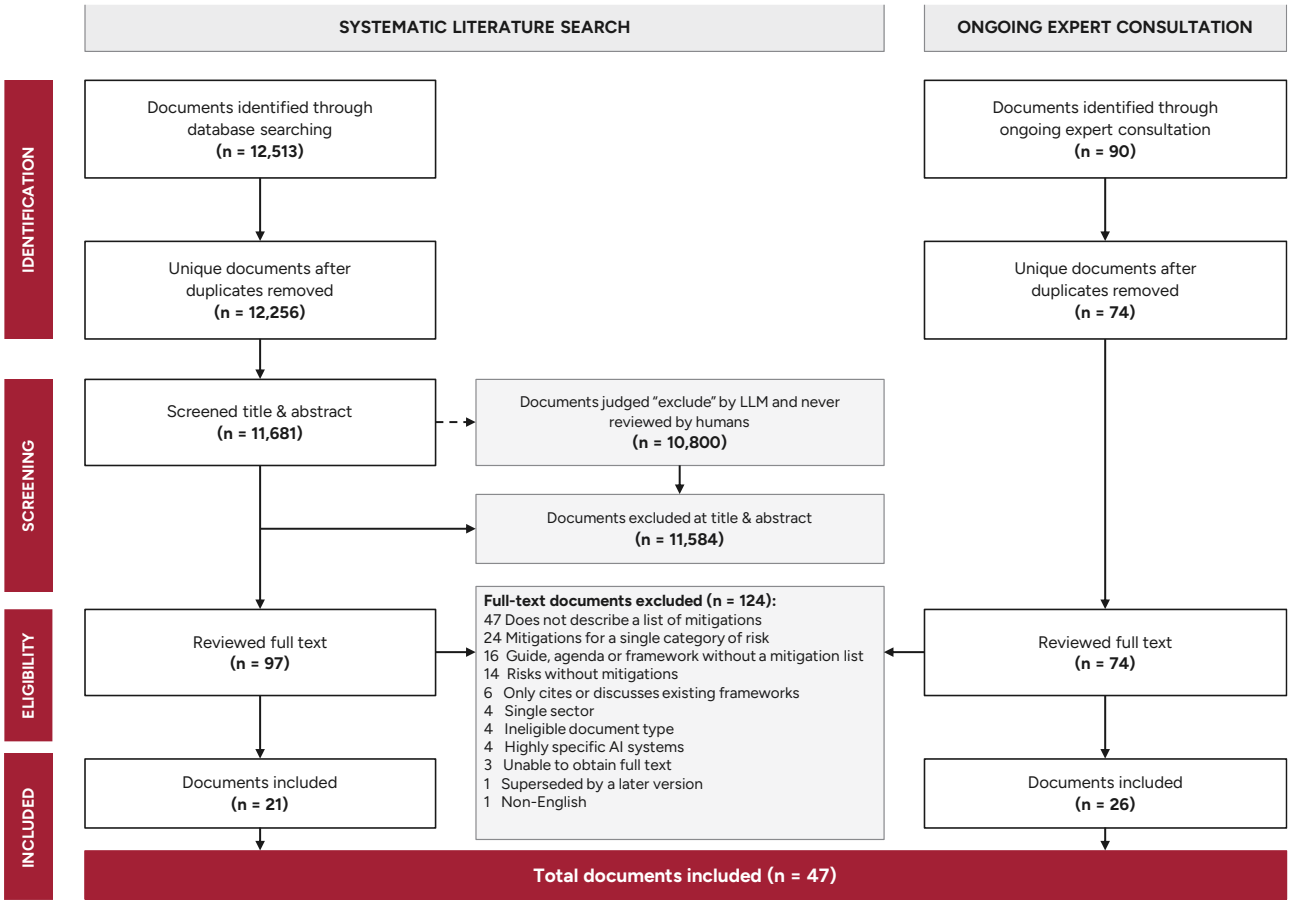
Methods

Overview of approach

To systematically identify mitigations for risks from AI and classify these mitigations against multiple taxonomies, we:

1. Conducted a systematic search of peer-reviewed and gray literature, and consulted with experts to identify additional candidate documents
2. Screened documents against inclusion and exclusion criteria at two stages: title and abstract screening, and subsequent full-text screening; we report on included documents at this step
3. Extracted mitigations from included documents
4. Developed a novel two-axis taxonomy of mitigation domains and instruments through iterative coding and expert consultation
5. Classified all mitigations against six dimensions across five taxonomies: mitigation domain, mitigation instrument, actor, lifecycle stage, risk addressed, and risk management function

Figure 14. PRISMA study flow diagram



Systematic search of peer-reviewed and gray literature

We conducted this study as a systematic review. Figure 14 summarizes the search and screening process as a PRISMA diagram, following established standards for transparent reporting in systematic reviews (Page et al., 2021).

Our search strategy included two stages: (1) conducting a systematic search of the peer-reviewed and gray literature to identify relevant documents, and (2) presenting a list of documents that had been included from stage 1 to experts in AI risks and mitigations, and requesting nomination of additional relevant documents.

We included reviews, articles, reports, and documents that proposed, presented, or substantially developed frameworks, taxonomies, lists, or other structured classifications of AI risk mitigations applied across multiple sectors and settings. We excluded book chapters, theses, commentaries, editorials, and protocols, based on previous searches that indicated including these documents would increase the number of search results without improving the number of relevant results. Documents were also excluded when they focused on (1) mitigations in a single location or sector (e.g., online learning platforms or medical imaging systems); (2) mitigations from highly specific AI systems, models, or applications (e.g., chatbots, image generation, autonomous vehicles); (3) mitigations for single categories of risks from AI (e.g., misinformation, privacy); (4) citing and discussing existing theories, frameworks, models, and taxonomies without new content; (5) research agendas, speculative discussion papers without clear empirical grounding, guides, procedures, or instructions for implementation of mitigations (without including the lists themselves), and risk management frameworks without a list of risk mitigations. Exclusion criterion (5) was added during data extraction to preserve data quality and comparability. Non-English documents were also excluded, and where a later edition of an included document was available, the earlier edition was excluded as superseded.

Search terms were generated by consulting studies consisting of a variety of search strategies (Eady et al., 2008; Lee et al., 2012). We chose terms that were shown empirically to balance sensitivity and specificity, using an initial “gold set” of 8 known-relevant documents that were included in an earlier non-systematic search of AI risk mitigations, with the resulting search string:

> TITLE-ABS-KEY (("artificial intelligence" OR ai) AND (framework* OR taxonom* OR classif* OR list) AND (mitigat* OR governance OR "risk management"))

Our systematic search queried an academic database, Scopus, to identify peer-reviewed publications, as well as two sources of gray literature, arXiv and Crossref, to identify indexed gray literature. All searches were conducted on 7 July 2025 with identified documents (n = 12,513) downloaded for screening. Deduplication removed 257 documents, for a full set of 12,256 documents.

Our expert consultation involved contacting 104 experts in AI risks and mitigations from our networks. We provided experts with the research protocol consisting of questions to guide the systematic review, inclusion and exclusion criteria, as well as the list of included and excluded documents at the full-text stage (to reduce duplicate recommendations). We asked experts to suggest additional documents for consideration; 90 documents were suggested at this step.

Document screening

Five trained reviewers carried out title and abstract and full-text screening. We used LLMs as assistive tools during screening, extraction, and classification, consistent with emerging practice in evidence synthesis (Cao et al., 2025; Nguyen-Trung et al., 2024).

In title and abstract screening, one author prepared the corpus of retrieved documents (n = 12,256) for screening and created a pipeline in which an LLM was presented with the title and abstract for each document using a structured prompt adapted from existing evidence synthesis research (Cao et al., 2025). Both an anonymized preview model on OpenRouter (horizon-alpha) and GPT-4.1 were used. A total of 575 documents were removed due to missing minimum information (e.g., title or abstract). Of the 11,681 documents that could be screened, the LLM returned a recommendation to include (n = 417), exclude (n = 10,961), or did not return a recommendation (treated as "uncertain"; n = 303). All "include" and "uncertain" documents, as well as a sample of 161 "exclude" documents, were screened in duplicate by human reviewers. Conflicts between pairs of human reviewers were resolved by a third human reviewer. Of these 881 documents, 97 were judged relevant and sought in full text.

In full-text screening of the documents obtained through systematic search, one author prepared the corpus of relevant documents (n = 97) for screening and created a pipeline in which Claude Sonnet 4 was presented with the title, abstract, and PDF full text for each document using a structured prompt. Three documents could not have their full text retrieved. Each full-text document was screened as a human-LLM pair, with conflicts resolved by a human reviewer who was blinded to the decision and source (e.g., whether the LLM recommended "include" and human recommended "exclude", or vice versa). After this process, 35 documents were retained for extraction. During extraction, 14 were subsequently excluded (13 under criterion 5 and 1 where the 2025 International AI Safety Report was superseded by the 2026 edition). In total, 21 documents were included from the systematic search.

In full-text screening of the 90 documents obtained through expert consultation, we first removed 16 duplicate or previously screened documents, then replicated the same pipeline as for the systematic search. After this process, 32 documents were retained for extraction. During extraction, 6 were subsequently excluded (3 under criterion 5 and 3 under existing criteria). In total, 26 documents were included from the expert consultation.

Included documents

From an initial search of 12,256 unique documents and 90 documents identified through expert consultation, we included 47 documents, presented in Table 9 (see also Supplementary Table A). Documents were frameworks, taxonomies, or structured lists of AI risk mitigations published between 2018 and 2026[7], with about half published in 2025. Documents were primarily weighted toward the gray literature; only 11 documents were peer reviewed. Documents were most likely to have first authors or organizations from the United States.

[7] Although our search was conducted in 2025, we included one 2026 document as it was an update of an earlier version (Bengio et al., 2026).

Together, these documents contributed a total of 2,206[8] AI risk mitigation records, with contributions varying from 3 to 258 mitigation records per document. The five documents with the largest number of individual mitigation records supplied 39% of all records in the database.

Extraction

The 47 included documents were extracted by one trained extractor and checked by a second trained extractor. For each document, the extractor located the document's structured list of AI risk mitigations and extracted each mitigation name, definition (if provided), additional information (e.g., examples, if provided), and structural information (e.g., if one mitigation was a subcategory of another mitigation). Extraction sought to faithfully reproduce the document authors' provided structure of the mitigations and to preserve the fidelity of their original categorizations and descriptions. This was an effort to capture the phenomenon directly rather than impose preconceived interpretations, in line with the principles of grounded theory (Corbin & Strauss, 2014). In addition, document metadata including the author(s), title, abstract, year of publication, document type, DOI and/or URL, and source or journal were extracted.

[8] Because our extraction preserved the hierarchies used by source documents, we excluded 413 parent-level records; only the lowest-level mitigation was classified to avoid double counting.

Table 9. Characteristics of included documents

Characteristic	Documents	% of all documents	Mitigation records	% of all records
Publisher sector				
Academic	15	32	843	38
Government or public body	12	26	708	32
Non-profit research institute	10	21	404	18
Industry	6	13	122	6
Standards organization	2	4	52	2
Individual or unaffiliated	2	4	77	3
Country or region				
United States	15	32	987	45
United Kingdom	11	23	423	19
China	7	15	317	14
European Union	6	13	183	8
Other	6	13	184	8
International	2	4	112	5
Document type				
Preprint or working paper	14	30	493	22
Organizational report	12	26	792	36
Peer-reviewed article	11	23	307	14
Government or regulatory guidance	5	11	449	20
Other gray literature	3	6	107	5
Technical standard or code of practice	2	4	58	3
Synthesizes existing work				
No	24	51	931	42
Yes	23	49	1,275	58
Publication year				
2018	1	2	20	1
2022	3	6	114	5
2023	5	11	354	16
2024	15	32	679	31
2025	22	47	963	44
2026	1	2	76	3
Total	***47***	***100***	***2,206***	***100***

Novel AI risk mitigation taxonomy development

Unable to identify a best-fit framework. We aimed to develop a mitigation taxonomy as an accessible and comprehensive structure for understanding the AI mitigation landscape, consistent with our previous development of AI risk taxonomies (Slattery et al., 2026) and as a replacement for a preliminary mitigation taxonomy based on a non-systematic search (Saeri et al., 2025)[9]. At first, we intended to use a best-fit framework approach, adapted from Slattery et al. (2026) and Carroll et al. (2011), to find a widely used and comprehensive existing framework or taxonomy of AI risk mitigations, and adapt it iteratively. However, after full-text review it became clear that the set of included documents was too diverse for any single document to serve as a best-fit framework. Existing frameworks were either too high-level (NIST, 2024), too narrow (e.g., AI system-specific), or combined multiple dimensions (lifecycle stage, actor, etc.) into a single set of categories.

Development of draft taxonomy centered on sociotechnical system. To initiate development of a novel mitigation taxonomy, two authors (AS, PS) reviewed all the extracted documents and the

[9] Note that this *preliminary* mitigation taxonomy, based on an evidence scan of 13 documents, is different from the *draft* mitigation taxonomy from the current systematic review that is mentioned later in the Methods.

database records, and proposed a hierarchical classification structure that organized mitigations by their position in the sociotechnical system: mitigations acting upon the *AI Systems*; mitigations acting upon *Organizations* that develop or deploy AI systems; and mitigations acting upon the wider *Ecosystems* that contain, regulate, or coordinate organizations and AI systems. Each top-level category would include subcategories (e.g., organizational security mitigations) and sub-subcategories (e.g., know-your-customer procedures). Taxonomy categories were developed through inductive-deductive iteration cycles consisting of specification of categories, classification of a sample of mitigation records, analysis of the mitigations, and revision of definitions (Nickerson et al., 2013). This produced a draft taxonomy consisting of 3 categories at the highest level (AI system, organization, ecosystem), 9 subcategories at the mid-level, and 35 categories at the lowest level.

Expert feedback on draft taxonomy. This draft taxonomy was shared with an expert panel of 97 AI risk and mitigation experts. We solicited feedback on the comprehensiveness and logic of the taxonomy (e.g., "Overall, how comprehensive is the structure of this taxonomy? How could the structure of this taxonomy be improved?") and the usefulness of the taxonomy (e.g., "Overall, how useful is this taxonomy as a shared framework for coordinating action on AI risks by policymakers, researchers, and practitioners?"). This led to 45 responses corresponding with 190 comments on improving the taxonomy and/or its presentation and accompanying documentation.

Feedback focused on three qualities of the taxonomy. First, experts were concerned about the conflation of locus (at the top-most level) and function (at lower levels) as nested within a single dimension. For example, *Research & Foundations*, which represented a function, was located within the *Organization* locus but could also cut across to the *Ecosystem* locus. Second, many comments identified a fragmentation problem, with the taxonomy's organizing principles splitting coherent concepts apart. Because lower levels were organized by overlapping mechanisms, single recognizable mitigations were split across categories: human-in-the-loop, for example, was distributed across policy, design standard, technical interface, and end-user practice. Third, experts were concerned about the complexity of the taxonomy, with 47 codes across a nested structure, which was difficult to remember and yielded lower inter-rater agreement. Additionally, expert feedback clustered around a lack of coverage in the taxonomy for certain mitigations, including (1) human-in-the-loop and agentic control, (2) end-user governance for both individuals and end-user organizations, (3) societal resilience mitigations, and (4) ex post liability and insurance regimes.

Taxonomy revision process. Based on this feedback, several candidate models of the refined taxonomy were developed that applied different sets of changes and different organizing principles based on the comments. Separate potential organizing principles emerged, including causal stage (when does the intervention act relative to harm), layer of defense (what protective barrier does it strengthen), system level (what level of the system is targeted), mechanism of influence (how does the intervention produce change), risk modified (which element of risk is altered by the mitigation), and functional capability (which capacity is improved). In deciding which organizing principles would structure the taxonomy, we considered our expert panel's feedback and drew widely from other frameworks in public policy, public health, cybersecurity, regulatory governance, and behavioral science that deal with multiple organizing principles and cross-cutting themes (Haddon, 1972; Hood, 1983; Howlett, 2009; Michie et al., 2011; Strom et al., 2020; World Health Organization, 2024).

Separation into mitigation domain and mitigation instrument taxonomies. The final version distinguished two dimensions of AI risk mitigations. First, the *mitigation domain* taxonomy organizes what the mitigation is intended to influence or improve: "what change occurs if this mitigation is successful". Examples include greater interpretability and transparency within the AI system, and greater accountability within organizations. These are further clustered into three orienting loci where the change occurs: *AI System*, *Human & Organization*, *Society & Ecosystem*. Second, the *mitigation instrument* taxonomy organizes the main mechanism of influence: "what is being done or produced to create change". These consist of more generic, cross-cutting levers of change that could be applied across multiple domains of AI risk mitigations (or other societal challenges), such as research, creation of technical artifacts, and laws and regulation.

LLM-assisted classification and validation

All records were classified against six dimensions across five taxonomies, representing different elements in organizing AI risk and mitigations.

1. **AI actor** (e.g., *Developer*, *Deployer*) that implements the mitigation (adapted from Australian Government, 2024; NIST, 2023)
2. **AI lifecycle stage** (e.g., *Plan & Design*, *Build & Use Model*) where the mitigation is implemented (adapted from NIST, 2023; OECD, 2022)
3. **AI risk** (e.g., *Discrimination & Toxicity*, *Privacy & Security*) that the mitigation is intended to address, coded to both the domain and subdomain levels (Slattery et al., 2026)
4. **AI risk management function** (e.g., *Govern*, *Map*, *Measure*, *Manage*) that the mitigation is aligned with, based on the NIST AI RMF functions (NIST, 2023; OECD, 2022)
5. **Mitigation domain** (e.g., *Interpretability & Transparency*, *Accountability & Responsibility*) that categorizes what the mitigation is intended to influence or improve
6. **Mitigation instrument** (e.g., *Research & Evidence*, *Law & Regulation*) that categorizes what the mitigation does or produces to address AI risk

We used a non-inferiority framework to determine whether an LLM could classify mitigations against the taxonomies with sufficient accuracy for analysis and reporting. Under this framework, we compared whether LLM-human agreement (LH) could reach non-inferiority compared to human-human agreement (HH) by varying the model and prompt. Non-inferiority was pre-specified as the lower bound of the 95% bootstrap confidence interval on Δ = mean(LH κ) − mean(HH κ) exceeding a margin of −0.10 κ.

Human classifications. We constructed a development and holdout set of human-rated mitigations to allow for prompt and model experimentation while having a clean dataset for final evaluation. A sample of 238 mitigation records stratified by source document was drawn from the database and was split into a 118-record development set and a 120-record holdout set. Five independent human raters blind to the development versus holdout split coded each record across all taxonomies. We calculated pairwise Cohen's κ to establish inter-rater reliability between all human-human (HH) pairs for each taxonomy.

LLM experimentation with classifications. For each taxonomy, we created an initial structured prompt that included fixed input (taxonomy definitions, classification guidance, examples of

categories) and variable input (mitigation name, description, and additional information), and required output in a structured XML format (primary code, secondary code, reasoning). This initial prompt was used with Gemini 3 Flash to classify all records in the development set. Non-inferiority was calculated on the development set. If the LLM-human agreement was inferior to the human-human agreement on the development set, we iterated by experimenting with different AI models (e.g., GPT-5.4) and prompts. For example, in some prompts we changed the fixed input to provide more guidance on easily confused categories; in others, we changed the variable input to provide the document's abstract or information about other mitigations from the document. The combination of model and prompt that performed best for each taxonomy against the development set was retained for testing against the holdout set.

Non-inferiority testing. We tested the best-performing LLM approach for each taxonomy against the holdout set and compared LLM-human agreement (LH) against human-human agreement (HH). In every case, LLM-human agreement was non-inferior to human-human agreement: AI actor (LH κ = 0.469, HH κ = 0.426, Δ = +0.043); lifecycle stage (LH κ = 0.517, HH κ = 0.372, Δ = +0.145); AI risk addressed (LH κ = 0.208, HH κ = 0.112, Δ = +0.096); RMF function (LH κ = 0.589, HH κ = 0.496, Δ = +0.093); mitigation domain (LH κ = 0.469, HH κ = 0.379, Δ = +0.090); and mitigation instrument (LH κ = 0.589, HH κ = 0.451, Δ = +0.138).

Database classification. After establishing that LLM classification of mitigations against the relevant taxonomies was non-inferior to human classification, we used the best-performing LLM approach to classify all database records against all taxonomies. In addition, all classifications are being human reviewed to confirm or correct the LLM classification. As of 8 October 2026 (corresponding with the version of the repository used for reporting in this paper), 100% of record categories have been classified. Of these, 49% have been LLM classified and 51% have been human reviewed or classified.

References

References marked with an asterisk (*) indicate the 47 documents included in the systematic review. The label in square brackets at the end of each of these references is the source identifier presented in tables.

Abdelwanis, M., Alarafati, H. K., Tammam, M. M. S., & Simsekler, M. C. E. (2024). Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *Journal of Safety Science and Resilience*, *5*(4), 460–469. https://doi.org/10.1016/j.jnlssr.2024.06.001

Actuarial Standards Board. (2014). *Risk evaluation in enterprise risk management* (ASOP No. 46). Actuarial Standards Board. https://www.actuarialstandardsboard.org/asops/risk-evaluation-enterprise-risk-management-2/

Alan Turing Institute. (2026). *AI Standards Hub*. https://aistandardshub.org/

Anthropic. (2026). *Responsible Scaling Policy (Version 3.0)*. Anthropic. https://www-cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf

Australian Government. (2024). *Voluntary AI Safety Standard*. https://www.industry.gov.au/sites/default/files/2024-09/voluntary-ai-safety-standard.pdf

Aven, T., Ben-Haim, Y., Andersen, H. B., Cox, T., Droguett, E. L., Greenberg, M., Guikema, S., Kröger, W., Renn, O., Thompson, K. M., & Zio, E. (2018). *Society for Risk Analysis Glossary*. Society for Risk Analysis. https://www.sra.org/wp-content/uploads/2020/04/SRA-Glossary-FINAL.pdf

*Barrett, A. M., Newman, J., Nonnecke, B., Madkour, N., Hendrycks, D., Murphy, E. R., Jackson, K., & Raman, D. (2025). *AI risk-management standards profile for General-Purpose AI (GPAI) and foundation models*. Center for Long-Term Cybersecurity, UC Berkeley. https://cltc.berkeley.edu/seeking-input-and-feedback-ai-risk-management-standards-profile-for-increasingly-multi-purpose-or-general-purpose-ai/ [Barrett2025]

*Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., … Mindermann, S. (2026). *International AI Safety Report 2026*. Department for Science, Innovation and Technology, AI Safety Institute. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026 [Bengio2026]

*Bernardi, J., Mukobi, G., Greaves, H., Heim, L., & Anderljung, M. (2025). *Societal adaptation to advanced AI* (arXiv:2405.10295). arXiv. https://doi.org/10.48550/arXiv.2405.10295 [Bernardi2025]

Bogen, M., Kerry, C. F., Kornbluh, K., Meltzer, J. P., Mir Teijeiro, M., Munkhbayar, E., Nadgir, N., Tabassi, E., Tanner, B., & Wirtschafter, V. (2026). *The HAIP Reporting Framework: Its value in global AI governance and recommendations for the future*. Brookings Institution. https://www.brookings.edu/articles/haip-reporting-framework-ai-governance/

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

*Buhl, M. D., Bucknall, B., & Masterson, T. (2025). *Emerging practices in frontier AI safety frameworks* (arXiv:2503.04746). arXiv. https://doi.org/10.48550/arXiv.2503.04746 [Buhl2025]

*Campos, S., Papadatos, H., Roger, F., Touzet, C., Quarks, O., & Murray, M. (2025). *A Frontier AI Risk Management Framework: Bridging the gap between current AI practices and established risk management* (arXiv:2502.06656). arXiv. https://doi.org/10.48550/arXiv.2502.06656 [Campos2025]

Cao, C., Arora, R., Cento, P., Budak, A., Manta, K., Farahani, E., Cecere, M., Selemon, A., Sang, J., Gong, L. X., Kloosterman, R., Jiang, S., Saleh, R., Margalik, D., Lin, J., Jomy, J., Xie, J., Chen, D., Gorla, J., … Bobrovitz, N. (2025). *Automation of systematic reviews with large language models*. medRxiv. https://doi.org/10.1101/2025.06.13.25329541

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*(1), 29. https://doi.org/10.1186/1471-2288-11-29

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*(9683), 86–89. https://doi.org/10.1016/s0140-6736(09)60329-9

*Chen, C., Gong, X., Liu, Z., Jiang, W., Goh, S. Q., & Lam, K.-Y. (2024). *AI safety landscape for Large Language Models: Taxonomy, state-of-the-art, and future directions* (arXiv:2408.12935). arXiv. https://doi.org/10.48550/arXiv.2408.12935 [Chen2024]

Cloud Security Alliance. (2026). *AI Controls Matrix (AICM) v1.1*. Cloud Security Alliance. https://cloudsecurityalliance.org/artifacts/ai-controls-matrix-v1-1

Corbin, J. M., & Strauss, A. L. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. SAGE.

*Crichton, K., Reddy, A., Ji, J., Crawford, A., Hoffmann, M., Shea-Blymyer, C., & Bansemer, J. (2025). *Harmonizing AI guidance: Distilling voluntary standards and best practices into a unified framework*. Center for Security and Emerging Technology. https://doi.org/10.51593/20240041 [Crichton2025]

*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* (arXiv:2401.05778). arXiv. https://doi.org/10.48550/arXiv.2401.05778 [Cui2024]

*Department for Science, Innovation and Technology. (2023, October 27). *Emerging processes for frontier AI safety*. https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety/emerging-processes-for-frontier-ai-safety [DSIT2023]

Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing language model deployment with risk cards*. arXiv. https://doi.org/10.48550/arXiv.2303.18190

Eady, A. M., Wilczynski, N. L., & Haynes, R. B. (2008). PsycINFO search strategies identified methodologically sound therapy studies and review articles for use by clinicians and researchers. *Journal of Clinical Epidemiology*, *61*(1), 34–40. https://doi.org/10.1016/j.jclinepi.2006.09.016

*Eisenberg, I. W., Gamboa, L., & Sherman, E. (2025). *The Unified Control Framework: Establishing a common foundation for enterprise AI governance, risk management and regulatory compliance* (arXiv:2503.05937). arXiv. https://doi.org/10.48550/arXiv.2503.05937 [Eisenberg2025]

*ETSI. (2024). *Securing artificial intelligence; Mitigation strategy report (ETSI TR 104 222 V1.2.1)*. https://www.etsi.org/deliver/etsi_tr/104200_104299/104222/01.02.01_60/tr_104222v010201p.pdf [ETSI2024]

*European Commission. (2025). *General-Purpose AI Code of Practice: Safety and security chapter*. European Commission. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai [AIOffice2025]

European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). Red-teaming for generative AI: Silver bullet or security theater? *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 421–437. https://doi.org/10.1609/aies.v7i1.31647

*Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5 [Floridi2018]

*Frontier Model Forum. (2025, June 30). *Frontier mitigations*. https://www.frontiermodelforum.org/technical-reports/frontier-mitigations/ [FMF2025]

*Gipiškis, R., San Joaquin, A., Chin, Z. S., Regenfuß, A., Gil, A., & Holtman, K. (2024). *Risk sources and risk management measures in support of standards for General-Purpose AI systems* (arXiv:2410.23472). arXiv. https://doi.org/10.48550/arXiv.2410.23472 [Gipiskis2024]

*Glasser, G., Cuéllar, M.-F., & O'Reilly, T. (2025, September). *Creating a standard for TAI governance*. https://www.lesswrong.com/posts/ntLxPrHvCShDnAPrH/creating-a-standard-for-tai-governance [Glasser2025]

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., Michie, S., Moher, D., & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, *383*(9913), 267–276. https://doi.org/10.1016/S0140-6736(13)62228-X

*Google. (2024). *Secure AI Framework (SAIF) Risk Map*. Google. https://saif.google/secure-ai-framework/controls [Google2024]

*Google. (2025). *Responsible AI progress report*. https://ai.google/static/documents/ai-responsibility-update-published-february-2025.pdf [Google2025]

Google DeepMind. (2026). *Frontier Safety Framework (Version 3.1)*. Google DeepMind. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/strengthening-our-frontier-safety-framework/frontier-safety-framework_3-1.pdf

Haddon, W. (1972). A logical framework for categorizing highway safety phenomena and activity. *The Journal of Trauma*, *12*(3), 193–207. https://doi.org/10.1097/00005373-197203000-00002

*Hodes, C., Salem, F., Corruble, V., Ségerie, C.-R., Claybrough, J., Veron, T., Majid, Z., Fan, J., & Lorin, A. (2025). *Global Risk and AI Safety Preparedness (GRASP)*. https://aisafetypreparedness.com/explore [Hodes2025]

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A.-W., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, *348*, g1687. https://doi.org/10.1136/bmj.g1687

Hood, C. C. (1983). *The tools of government*. Macmillan.

Hopewell, S., Chan, A.-W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., Tunn, R., Aggarwal, R., Berkwits, M., Berlin, J. A., Bhandari, N., Butcher, N. J., Campbell, M. K., Chidebe, R. C. W., Elbourne, D., Farmer, A., Fergusson, D. A., Golub, R. M., Goodman, S. N., … Boutron, I. (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ*, *389*, e081123. https://doi.org/10.1136/bmj-2024-081123

Howlett, M. (2009). Governance modes, policy regimes and operational plans: A multi-level nested model of policy instrument choice and policy design. *Policy Sciences*, *42*(1), 73–89. https://doi.org/10.1007/s11077-009-9079-1

ISO. (2018). *Risk management—Guidelines* (ISO 31000:2018). ISO. https://www.iso.org/standard/65694.html

ISO. (2022). *Risk management—Vocabulary* (ISO 31073:2022). ISO. https://www.iso.org/standard/79637.html

ISO/IEC. (2023). *Information technology—Artificial intelligence—Management system* (ISO/IEC 42001:2023). ISO. https://www.iso.org/standard/81230.html

*Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Vierling, L., Hong, D., Zhou, J., Zhang, Z., Zeng, F., Dai, J., Pan, X., Ng, K. Y., O'Gara, A., Xu, H., Tse, B., … Gao, W. (2023). *AI alignment: A comprehensive survey* (arXiv:2310.19852). arXiv. https://doi.org/10.48550/arXiv.2310.19852 [Ji2023]

*Jones, A. (2024, August 10). *The AI regulator's toolbox: A list of concrete AI governance practices*. https://adamjones.me/blog/ai-regulator-toolbox/ [Jones2024]

*Joshi, H., Hassani, S., Gandhi, D., & Hartman, L. (2025). *Approaches to responsible governance of GenAI in organizations* (arXiv:2504.17044). arXiv. https://doi.org/10.48550/arXiv.2504.17044 [Joshi2025]

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2). https://doi.org/10.14763/2020.2.1469

Lee, E., Dobbins, M., DeCorby, K., McRae, L., Tirilis, D., & Husson, H. (2012). An optimal search filter for retrieving systematic reviews and meta-analyses. *BMC Medical Research Methodology*, *12*(1), 51. https://doi.org/10.1186/1471-2288-12-51

*Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI Pattern Catalogue: A collection of best practices for AI governance and engineering. *ACM Computing Surveys*, *56*(7), 1–35. https://doi.org/10.1145/3626234 [Lu2024]

*Ma, X., Gao, Y., Wang, Y., Wang, R., Wang, X., Sun, Y., Ding, Y., Xu, H., Chen, Y., Zhao, Y., Huang, H., Li, Y., Wu, Y., Zhang, J., Zheng, X., Bai, Y., Wu, Z., Qiu, X., Zhang, J., … Jiang, Y.-G. (2025). *Safety at scale: A comprehensive survey of large model and agent safety* (arXiv:2502.05206). arXiv. https://doi.org/10.48550/arXiv.2502.05206 [Ma2025]

*Maas, M. M. (2023). *Advanced AI governance: A literature review of problems, options, and proposals*. SSRN. https://doi.org/10.2139/ssrn.4629460 [Maas2023]

*Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). *Putting AI ethics into practice: The hourglass model of organizational AI governance* (arXiv:2206.00335). arXiv. https://doi.org/10.48550/arXiv.2206.00335 [Mantymaki2022]

Mengesha, I. (2026). *The coordination gap in frontier AI safety policies* (arXiv:2603.10015; Version 2). arXiv. https://doi.org/10.48550/arXiv.2603.10015

Michie, S., Fixsen, D., Grimshaw, J. M., & Eccles, M. P. (2009). Specifying and reporting complex behaviour change interventions: The need for a scientific method. *Implementation Science*, *4*, 40. https://doi.org/10.1186/1748-5908-4-40

Michie, S., van Stralen, M. M., & West, R. (2011). The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science*, *6*(1), 42. https://doi.org/10.1186/1748-5908-6-42

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for developers of health research reporting guidelines. *PLoS Medicine*, *7*(2), e1000217. https://doi.org/10.1371/journal.pmed.1000217

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

*National Technical Committee 260 on Cybersecurity of SAC. (2024, September). *AI Safety Governance Framework (Version 1.0)*. https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf [NTC2602024]

Nguyen-Trung, K., Saeri, A. K., & Kaufman, S. (2024). Applying ChatGPT and AI-powered tools to accelerate evidence reviews. *Human Behavior and Emerging Technologies*, *2024*(1), 8815424. https://doi.org/10.1155/2024/8815424

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

NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/nist.ai.100-1

*NIST. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST. https://doi.org/10.6028/NIST.AI.600-1 [NIST2024]

OECD. (2022). *OECD Framework for the Classification of AI Systems* (OECD Digital Economy Papers No. 323). https://doi.org/10.1787/cb6d9eca-en

OECD. (2023). *Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI* (OECD Digital Economy Papers No. 349). https://doi.org/10.1787/2448f04b-en

OECD. (2025). *Hiroshima AI Process (HAIP) Reporting Framework*. Organisation for Economic Co-operation and Development. https://transparency.oecd.ai/

OECD.AI Policy Observatory. (2026a). *Catalogue of Tools & Metrics for Trustworthy AI*. Organisation for Economic Co-operation and Development (OECD). https://oecd.ai/en/catalogue/tools

OECD.AI Policy Observatory. (2026b). *OECD AI Incidents and Hazards Monitor (AIM)*. https://oecd.ai/en/incidents

OECD.AI Policy Observatory. (2026c). *OECD.AI Policy Navigator: Database of national AI policy initiatives*. Organisation for Economic Co-operation and Development (OECD). https://oecd.ai/en/dashboards/policy-initiatives

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

*Pathan, M. K., & Shah, A. (2025). Ethical considerations and responsible governance of generative AI: A systematic review. *Premier Journal of Artificial Intelligence*. https://doi.org/10.70389/pjai.100016 [Pathan2025]

Perset, K., & Fialho Esposito, S. (2025). *How are AI developers managing risks? Insights from responses to the reporting framework of the Hiroshima AI Process Code of Conduct* (OECD Artificial Intelligence Papers). OECD Publishing. https://doi.org/10.1787/658c2ad6-en

Proctor, E. K., Powell, B. J., & McMillen, J. C. (2013). Implementation strategies: Recommendations for specifying and reporting. *Implementation Science*, *8*(1), 139. https://doi.org/10.1186/1748-5908-8-139

*Radanliev, P. (2025). Frontier AI regulation: What form should it take? *Frontiers in Political Science*, *7*. https://doi.org/10.3389/fpos.2025.1561776 [Radanliev2025]

Ramchurn, G., Lloyd Owen, A., Ao, S., Barnard, P., Gomez Bergin, A., Parisio, I., Valoor, A., Waheed, M., Ten Holter, C., Portillo, V., Procter, R., Batterham, P., Downer, J., Winter, P., Krook, J., Blockx, J., Braunschweig, B., Quintero, K., Poretschkin, M., … Gillings, E. (2025). *Frameworks and toolkits for assuring responsible AI* (RAi/002). Responsible AI UK. https://doi.org/10.5258/RAi/002

Responsible AI Collaborative. (2026). *AI Incident Database*. https://incidentdatabase.ai/

Saeri, A. K., George, S. L., Graham, J., Lacarriere, C. D., Slattery, P., Noetel, M., & Thompson, N. (2025). *Mapping AI risk mitigations: Evidence scan and preliminary AI Risk Mitigation Taxonomy* (arXiv:2512.11931). arXiv. https://doi.org/10.48550/arXiv.2512.11931

Saeri, A. K., Graham, J., Noetel, M., Slattery, P., Ah-king, D., Aittokallio, E., Akindehin, I., Mahdi, A. A., Alhajjar, E., Lipcsey, R. A., Ang, G., Azam, C. M., Azaria, A., Balkissoon, R., Barberá, I., Bareato, C., Barry, J., Basehart, M., Bean, A. M., … Thompson, N. (2026). *Prioritization of risks from artificial intelligence: A Delphi study of 272 international experts* (arXiv:2606.04490). arXiv. https://doi.org/10.48550/arXiv.2606.04490

*Sakib, M. N., Islam, M. A., Pathak, R., & Arifin, M. M. (2024). *Risks, causes, and mitigations of widespread deployments of large language models (LLMs): A survey*. 1–7. https://doi.org/10.1109/AIBThings63359.2024.10863356 [Sakib2024]

Schiff, D. S. (2026). Strategies for harmonizing fragmented AI ethics frameworks, standards, and regulations. In W. Xu (Ed.), *Handbook of human-centered artificial intelligence* (pp. 1–45). Springer. https://doi.org/10.1007/978-981-95-3658-0_82

Schiff, D. S., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the principles to practices gap in AI. *IEEE Technology and Society Magazine*, *40*(2), 81–94. https://doi.org/10.1109/MTS.2021.3056286

*Schuett, J., Dreksler, N., Anderljung, M., McCaffary, D., Heim, L., Bluemke, E., & Garfinkel, B. (2023). *Towards best practices in AGI safety and governance: A survey of expert opinion* (arXiv:2305.07153). arXiv. https://doi.org/10.48550/arXiv.2305.07153 [Schuett2023]

*Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., Jain, R., Greig, R., Albanie, S., Emmons, S., Farquhar, S., Krier, S., Rajamanoharan, S., Bridgers, S., Ijitoye, T., … Dragan, A. (2025). *An approach to technical AGI safety and security* (arXiv:2504.01849). arXiv. https://doi.org/10.48550/arXiv.2504.01849 [Shah2025]

*Shi, D., Shen, T., Huang, Y., Li, Z., Leng, Y., Jin, R., Liu, C., Wu, X., Guo, Z., Yu, L., Shi, L., Jiang, B., & Xiong, D. (2024). *Large Language Model safety: A holistic survey* (arXiv:2412.17686). arXiv. https://doi.org/10.48550/arXiv.2412.17686 [Shi2024]

*Sigfrids, A., Nieminen, M., Leikas, J., & Pikkuaho, P. (2022). How should public administrations foster the ethical development and use of artificial intelligence? A review of proposals for developing governance of AI. *Frontiers in Human Dynamics*, *4*, 858108. https://doi.org/10.3389/fhumd.2022.858108 [Sigfrids2022]

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2026). The AI risk repository: A meta-review, database, and taxonomy of risks from artificial intelligence. *Patterns*, *7*(5), 101517. https://doi.org/10.1016/j.patter.2026.101517

*Somani, E., Friedman, A., Wu, H., Lu, M., Byrd, C., van Soest, H., & Zakaria, S. (2025). *Strengthening emergency preparedness and response for AI loss of control incidents*. RAND Corporation. https://doi.org/10.7249/RRA3847-1 [Somani2025]

*Srikumar, M., Chang, J., & Chmielinski, K. (2024). *Risk mitigation strategies for the open foundation model value chain*. Partnership on AI. https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/ [PAI2024]

*Srivastava, A., & Panda, S. (2024). *A formal framework for assessing and mitigating emergent security risks in generative AI models: Bridging theory and dynamic risk mitigation* (arXiv:2410.13897). arXiv. https://doi.org/10.48550/arXiv.2410.13897 [Srivastava2024]

Strom, B. E., Applebaum, A., Miller, D. P., Nickels, K. C., Pennington, A. G., & Thomas, C. B. (2020). *MITRE ATT&CK: Design and philosophy* (MP180360R1). The MITRE Corporation. https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf

Tanner, B., Belle, D., Tabassi, E., Kerry, C. F., Kyosovska, N., Renda, A., & Wyckoff, A. W. (2026, July 17). *New tools for mapping a fragmented AI governance landscape*. Brookings. https://www.brookings.edu/articles/new-tools-for-mapping-a-fragmented-ai-governance-landscape/

*Tkeshelashvili, M., & Saade, T. (2025). *Navigating AI compliance, part 2: Risk mitigation strategies for safeguarding against future failures*. Institute for Security and Technology. https://securityandtechnology.org/wp-content/uploads/2025/05/Navigating-AI-Compliance-Part-2-Risk-Mitigation-Strategies-for-Safeguarding-Against-Future-Failures.pdf [Tkeshelashvili2025]

*Tong, R. J., Cortês, M., DeFalco, J. A., Underwood, M., & Zalewski, J. (2025). *A first-principles based risk assessment framework and the IEEE P3396 standard*. 1588–1595. https://doi.org/10.1109/CAI64502.2025.00237 [Tong2025]

*Tse, B., Fang, L., Xu, J., Duan, Y., & Shao, J. (2025). *Frontier AI Risk Management Framework (v1.0)*. Concordia AI. https://concordia-ai.com/wp-content/uploads/2025/07/Frontier-AI-Risk-Management-Framework-v1.0-en.pdf [Tse2025]

*UK National Cyber Security Centre (NCSC). (2023). *Guidelines for secure AI system development*. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf [NCSC2023]

United Nations. (2026). *Preliminary report of the Independent International Scientific Panel on AI: Evidence-based assessment of opportunities, risks and impacts of artificial intelligence*. https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report

*US National Security Agency's Artificial Intelligence Security Center (AISC). (2024, April 15). *Deploying AI systems securely*. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/deploying-ai-systems-securely [AISC2024]

US-EU Trade and Technology Council. (2022). *TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management*. https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management

*Uuk, R., Brouwer, A., Schreier, T., Dreksler, N., Pulignano, V., & Bommasani, R. (2024). *Effective mitigations for systemic risks from General-Purpose AI* (arXiv:2412.02145). arXiv. https://doi.org/10.48550/arXiv.2412.02145 [Uuk2024]

*Wang, K., Zhang, G., Zhou, Z., Wu, J., Yu, M., Zhao, S., Yin, C., Fu, J., Yan, Y., Luo, H., Lin, L., Xu, Z., Lu, H., Cao, X., Zhou, X., Jin, W., Meng, F., Xu, S., Mao, J., … Liu, Y. (2025). *A comprehensive survey in LLM (-agent) full stack safety: Data, training and deployment* (arXiv:2504.15585). arXiv. https://doi.org/10.48550/arXiv.2504.15585 [Wang2025]

*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 [Wirtz2022]

World Health Organization. (2024). *Compendium of WHO and other UN guidance on health and environment: Version with International Classification of Health Intervention (ICHI) codes*. World Health Organization. https://www.who.int/publications/i/item/9789240088061

*Zhang, Z., Lei, L., Yang, J., Huang, X., Lu, Y., Cui, S., Chen, R., Zhang, Q., Wang, X., Wang, H., Li, H., Lei, X., Pan, C., Sha, L., Wang, H., & Huang, M. (2025). *AISafetyLab: A comprehensive framework for AI safety evaluation and improvement* (arXiv:2502.16776). arXiv. https://doi.org/10.48550/arXiv.2502.16776 [Zhang2025]

Author affiliations and ORCIDs

	Author	Affiliation	ORCID
1	Alexander K Saeri	MIT FutureTech School of Psychology, The University of Queensland	0000-0001-9254-0300
2	Kun Zhao	MIT FutureTech School of Psychology, The University of Queensland	0000-0002-4947-2418
3	Peter Slattery	MIT FutureTech	0000-0002-6083-378X
4	Michael Noetel	School of Psychology, The University of Queensland MIT FutureTech	0000-0002-6563-8203
5	Jess Graham	MIT FutureTech School of Psychology, The University of Queensland	0009-0003-9650-2792
6	Aileen Liang	Wellesley College	0009-0008-7440-3888
7	Catherine Simons	Cambridge Boston Alignment Initiative	
8	Elisa Zhang	MIT	0009-0005-7702-0733
9	Fatimeh Al Ghannam	MIT	
10	Soniya Agrawal	UC Berkeley	
11	Spencer Michaels	Cambridge Boston Alignment Initiative	0009-0003-0828-5637
12	Edla Aittokallio	Independent	
13	Gary Ang	Independent	0000-0001-5922-8956
14	Keita Azuma	United Nations University University of Cambridge	
15	Isabel Barberá	Independent	0009-0006-7508-5629
16	Peter Barnett	Machine Intelligence Research Institute	0000-0002-6100-7607
17	Otto Barten	Existential Risk Observatory	
18	Samantha Augusta Bennett	Stanford University	0009-0007-8499-6835
19	Jan Brauner	EU AI Office	0000-0002-1588-5724
20	Tao Burga	Institute for Progress	0009-0005-2698-4997
21	Ze Shen Chin	AI Standards Lab Oxford Martin AI Governance Initiative	0009-0006-9842-5681
22	Jan Llenzl Dagohoy	Independent	0000-0002-2357-8476
23	Elizabeth M Daly	IBM Research	0000-0003-0162-3683
24	Tom David	GPAI Policy Lab Oxford Martin AI Governance Initiative	0009-0009-1178-1950
25	Yawen Duan	Safe AI Forum	0000-0002-5124-1192
26	Ian W Eisenberg	Credo AI	0000-0002-4038-071X
27	Rokas Gipiškis	AI Standards Lab Vilnius University	0000-0001-5166-0920
28	Emily Grundy	Good Ancestors	0000-0002-4059-0311
29	Hans Gundlach	MIT	0000-0001-5499-5072
30	Adam Jones	Independent	0009-0007-5138-9633
31	Himanshu Joshi	Vector Institute for Artificial Intelligence University of Texas, Austin	0000-0002-9950-4625
32	Arturs Kanepajs	Independent	0009-0005-9880-1993

	Author	Affiliation	ORCID
33	Noam Kolt	Hebrew University	0009-0007-2538-6295
34	Sophia Lloyd George	Brown University	
35	Jayson Lynch	MIT FutureTech	0000-0003-0801-1671
36	Nada Madkour	UC Berkeley, Center for Long-Term Cybersecurity, AI Security Initiative	0009-0002-9988-9008
37	Claudio Mayrink Verdun	MIT FutureTech Harvard University	0000-0003-2079-797X
38	Malcolm Murray	SaferAI	0009-0001-4676-7413
39	Simon Mylius	MIT FutureTech Arcola AI	0009-0006-6205-1278
40	Kwan Yee Ng	Concordia AI	0009-0009-1794-2882
41	Ben Norman	Independent	
42	Max Reddel	Arq Foundation	0000-0002-0409-9365
43	Graham H Ryan	Independent	0009-0003-5883-5027
44	Ayrton San Joaquin	AI Standards Lab	0000-0002-4282-9170
45	Hugo Save	Existential Risk Observatory	0009-0006-6551-4282
46	Daniel S Schiff	Purdue University Governance and Responsible AI Lab	0000-0002-4376-7303
47	Yuanyuan Sun	Open Community for AI Safety China	0009-0009-8811-3424
48	Mariami Tkeshelashvili	The Institute for Security and Technology	
49	Cristian Trout	Artificial Intelligence Underwriting Company	0009-0000-9958-5077
50	Brian Tse	Concordia AI	0009-0001-8639-4287
51	Risto Uuk	Future of Life Institute KU Leuven	0009-0001-1665-3976
52	Lucia Velasco	Maastricht University Oxford Martin School AI Governance Initiative	0000-0001-9534-6343
53	Gabriel Weil	University of Houston Law Center Institute for Law and AI	
54	Neil Thompson	MIT FutureTech	0000-0001-9888-573X

Supplementary tables

Supplementary Table A. Characteristics of included documents, by document

Title	Source	Publisher sector	Country or region	Document type	Review	Mitigation records
EU General-Purpose AI Code of Practice; Safety and Security Section	AIOffice2025	Government or public body	European Union	Technical standard or code of practice	No	32
Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems	AISC2024	Government or public body	United States	Government or regulatory guidance	No	15
AI Risk-Management Standards Profile for General-Purpose AI (GPAI) and Foundation Models	Barrett2025	Academic	United States	Organizational report	No	125
International AI Safety Report	Bengio2026	Government or public body	United Kingdom	Organizational report	Yes	76
Societal Adaptation to Advanced AI	Bernardi2025	Non-profit research institute	United Kingdom	Preprint or working paper	No	3
Emerging Practices in Frontier AI Safety Frameworks	Buhl2025	Government or public body	United Kingdom	Organizational report	Yes	11
A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management	Campos2025	Non-profit research institute	European Union	Preprint or working paper	No	11
Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations	Chen2024	Academic	Other	Preprint or working paper	Yes	15
Harmonizing AI Guidance: Distilling Voluntary Standards and Best Practices into a Unified Framework	Crichton2025	Academic	United States	Organizational report	Yes	258
Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems	Cui2024	Government or public body	China	Preprint or working paper	Yes	25
Emerging processes for frontier AI safety	DSIT2023	Government or public body	United Kingdom	Government or regulatory guidance	Yes	138
The Unified Control Framework: Establishing a Common Foundation for Enterprise AI Governance, Risk Management and Regulatory Compliance	Eisenberg2025	Industry	United States	Preprint or working paper	No	42
Securing Artificial Intelligence; Mitigation Strategy Report	ETSI2024	Standards organization	European Union	Technical standard or code of practice	Yes	26
AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	Floridi2018	Academic	United Kingdom	Peer-reviewed article	No	20
Frontier Mitigations	FMF2025	Industry	United States	Organizational report	Yes	21
Risk Sources and Risk Management Measures in Support of Standards for General-Purpose AI Systems	Gipiskis2024	Non-profit research institute	International	Preprint or working paper	Yes	82
Creating a Standard for TAI Governance	Glasser2025	Individual or unaffiliated	United States	Other gray literature	Yes	35
Google Secure AI Framework	Google2024	Industry	United States	Organizational report	No	24
Responsible AI Progress Report	Google2025	Industry	United States	Organizational report	No	11
Global Risk and AI Safety Preparedness (GRASP)	Hodes2025	Academic	International	Other gray literature	Yes	30

Title	Source	Publisher sector	Country or region	Document type	Review	Mitigation records
AI Alignment: A Comprehensive Survey	Ji2023	Academic	China	Peer-reviewed article	Yes	21
The AI regulator's toolbox: A list of concrete AI governance practices	Jones2024	Individual or unaffiliated	United Kingdom	Other gray literature	Yes	42
Approaches to Responsible Governance of GenAI in Organizations	Joshi2025	Non-profit research institute	Other	Peer-reviewed article	No	11
Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering	Lu2024	Government or public body	Other	Peer-reviewed article	Yes	55
Safety at Scale: A Comprehensive Survey of Large Model and Agent Safety	Ma2025	Academic	China	Peer-reviewed article	Yes	85
Advanced AI Governance: A Literature Review of Problems, Options, and Proposals	Maas2023	Non-profit research institute	United States	Organizational report	Yes	128
Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance	Mantymaki2022	Academic	European Union	Preprint or working paper	No	67
Guidelines for secure AI development	NCSC2023	Government or public body	United Kingdom	Government or regulatory guidance	No	17
Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)	NIST2024	Government or public body	United States	Government or regulatory guidance	No	212
AI Safety Governance Framework	NTC2602024	Government or public body	China	Government or regulatory guidance	No	67
Risk Mitigation Strategies for the Open Foundation Model Value Chain	PAI2024	Non-profit research institute	United States	Organizational report	No	15
Ethical Considerations and Responsible Governance of Generative AI: A Systematic Review	Pathan2025	Industry	Other	Peer-reviewed article	Yes	16
Frontier AI regulation: what form should it take?	Radanliev2025	Academic	United Kingdom	Peer-reviewed article	Yes	19
Risks, Causes, and Mitigations of Widespread Deployments of Large Language Models (LLMs): A Survey	Sakib2024	Academic	Other	Peer-reviewed article	Yes	7
Towards best practices in AGI safety and governance: A survey of expert opinion	Schuett2023	Non-profit research institute	United Kingdom	Preprint or working paper	No	50
An Approach to Technical AGI Safety and Security	Shah2025	Industry	United Kingdom	Preprint or working paper	No	8
Large Language Model Safety: A Holistic Survey	Shi2024	Academic	China	Preprint or working paper	Yes	58
How Should Public Administrations Foster the Ethical Development and Use of Artificial Intelligence? A Review of Proposals for Developing Governance of AI	Sigfrids2022	Government or public body	European Union	Peer-reviewed article	Yes	15
Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents	Somani2025	Non-profit research institute	United Kingdom	Organizational report	No	39
A Formal Framework for Assessing and Mitigating Emergent Security Risks in Generative AI Models: Bridging Theory and Dynamic Risk Mitigation	Srivastava2024	Academic	United States	Preprint or working paper	No	10
Navigating AI Compliance, Part 2: Risk Mitigation Strategies for	Tkeshelashvili2025	Non-profit research institute	United States	Organizational report	No	39

Title	Source	Publisher sector	Country or region	Document type	Review	Mitigation records
Safeguarding Against Future Failures						
A First-Principles Based Risk Assessment Framework and the IEEE P3396 Standard	Tong2025	Standards organization	United States	Peer-reviewed article	No	26
Frontier AI Risk Management Framework (v1.0)	Tse2025	Government or public body	China	Organizational report	No	45
Effective Mitigations for Systemic Risks from General-Purpose AI	Uuk2024	Non-profit research institute	United States	Preprint or working paper	No	26
A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment	Wang2025	Academic	Other	Preprint or working paper	Yes	80
Governance of artificial intelligence: A risk and guideline-based integrative framework	Wirtz2022	Academic	European Union	Peer-reviewed article	Yes	32
AISafetyLab: A Comprehensive Framework for AI Safety Evaluation and Improvement	Zhang2025	Academic	China	Preprint or working paper	No	16

Note. All 47 included documents, ordered alphabetically by source. "Records" is the number of eligible mitigation records extracted from that document.

Supplementary Table B. Example mitigation records for each mitigation domain

Locus / Domain	Example mitigation records	Source
AI System		
Interpretability & Transparency	Thought Process and Internal State Monitoring	Wang2025
	Explainability and Interpretability	Crichton2025
	Develop framework to enhance explainability	Floridi2018
Alignment & Value Learning	RL-based Alignment	Zhang2025
	Behaviour/model specification	Bengio2026
	LLM debiasing techniques	Wang2025
Capability Control & Containment	Unlearning	Zhang2025
	Capability restrictions	Uuk2024
	Restrict web access during AI training	Gipiskis2024
Robustness & Adversarial Resistance	Red-teaming	Jones2024
	Adversarial Robustness	Uuk2024
	Robustness certificates	Gipiskis2024
Operational Safety & Control	Input and output filtering	Uuk2024
	Develop and test model shutdown measures	Somani2025
	Anomaly detection	Wang2025
Privacy & Data Protection	Quantifying privacy risks of AI models	Gipiskis2024
	Privacy Program	Crichton2025
	Consider the principle of data minimisation where possible	DSIT2023
Security & Access Control	KYC screenings	Uuk2024
	Enhance network and supply chain security	NTC2602024
	Applying cybersecurity standards	Buhl2025
Content Authenticity & Provenance	AI-content watermarking	Jones2024
	Develop a traceability management system for AI services	NTC2602024
	Implement Disclosure Mechanisms for AI-generated Content	PAI2024
Human Oversight & Agentic Control	Human in the loop	Jones2024
	Agent identifier	Tse2025
	Agent User Control	Google2024
Human & Organization		
Training & Capacity-building	Raise staff awareness of threats and risks	NCSC2023
	Conduct regular training and scenario drills	Somani2025
	Emergency response drills	Tse2025
Accountability & Responsibility	Third-party auditing	Jones2024
	AI safety and security committee	Tse2025
	Safety incident reports and security information sharing	Uuk2024
Integrity & Ethics	Whistleblower protections	Uuk2024
	Ethics committee	Lu2024
	Institutional Review Boards (IRBs)	Maas2023
Risk Management & Assurance	Safety cases	Jones2024
	Pre-deployment risk assessments	Uuk2024
	Risk register	Tse2025
Society & Ecosystem		
End-use Governance	"Opt out" option for end-users	Tkeshelashvili2025
	Avoid complete reliance on AI	NTC2602024
	User Transparency and Controls	Google2024
Compute, Hardware, & Resource Stewardship	Compute governance	Maas2023
	Hardware-based monitoring mechanisms	Bengio2026
	Pre-registration of large training runs	Schuett2023
Ecosystem Coordination	Initiate multilateral treaty negotiations	Maas2023
	Standards development	Jones2024
	Coordinate with developers and national authorities	Somani2025
Societal Resilience & Recovery	Incident databases	Maas2023
	Societal adaptation	Jones2024
	Crisis Management Plans	Glasser2025

Note. Three example mitigation records for each of the 17 mitigation domains, chosen as clear-cut members of the domain and drawn from three different source documents. Mitigation names are as given in the source document. "Other" categories are not shown.

Supplementary Table C. Example mitigation records for each mitigation instrument

Function / Instrument	Example mitigation records	Source
Inform		
Research & Evidence	Advance research on containment and layered defences	Somani2025
	Toxicity Datasets	Ma2025
	Forecasting	Glasser2025
Information Sharing & Disclosure	Sharing safety cases	Uuk2024
	Collect and share lessons learned	NCSC2023
	Share intelligence with relevant national security stakeholders	Somani2025
Govern		
Law & Regulation	Licensing	Jones2024
	Regulate Open Sourcing of AI Models with Dangerous Capabilities	Hodes2025
	Build adequate legal AI frameworks	Wirtz2022
External Standards	Voluntary Commitments	Glasser2025
	Security standards	Schuett2023
	Develop and Adhere to Transparency Reporting Standards	PAI2024
Internal Policies	Establish Decommissioning and Incident Response Policies	PAI2024
	Establish frontier AI safety framework	Eisenberg2025
	Access Policy	Crichton2025
Implement		
Technical Artifacts	Content classifiers	Jones2024
	Unlearning Techniques	Uuk2024
	Synthesized Sequence Filtering	Hodes2025
Practices & Processes	Evaluation and approval process for major decisions	Tse2025
	Review incident and shutdown procedures	Somani2025
	Securing internal change management processes	Jones2024
Monitoring	Abuse monitoring	Jones2024
	Real-time monitoring	Wang2025
	Monitor Misuses, Unintended Uses, and User Feedback	PAI2024
Evaluation & Audit	Safety-purpose benchmarks	Wang2025
	Third-party audits mechanisms	Tse2025
	Adversarial Testing Module	Srivastava2024

Note. Three example mitigation records for each of the 9 mitigation instruments, chosen as clear-cut members of the instrument and drawn from three different source documents. Mitigation names are as given in the source document. "Other" categories are not shown.

Supplementary Table D. Mitigation records that could not be classified to a mitigation domain

Mitigation record	Source
Consider principles relevant to foundation models and advanced AI	Barrett2025
Defence in-depth	Bengio2026
Forecasting	Bengio2026
Implementing comprehensive mitigations:	Buhl2025
Conduct research, collaborating with external stakeholders and organisations as needed, to identify and mitigate the risks and limitations of AI	DSIT2023
Release strategy disagreement between developers	Gipiskis2024
AI for Enhanced Efficiency and Capability	Glasser2025
Forecasting	Glasser2025
Robust compliance controls	Joshi2025
Blockchain smart contracts	Maas2023
Institutional, software, and hardware mechanisms for enabling developers to make verifiable claims	Maas2023
Online information repositories	Maas2023
Provide actionable technical information	Maas2023
Pursue specific career paths with key actors to contribute to good policymaking	Maas2023
Safety guidelines for AI development and application	NTC2602024
Continuous Monitoring	Pathan2025
AI cryptography and National Security Risks	Radanliev2025
Discussion on bias mitigation and digital inequality	Radanliev2025
Ethics and Compliance Proposals	Shi2024
Toward Effective and Unified Safety Mechanisms	Shi2024
Operationalization mechanisms	Sigfrids2022
Update auditing, evaluation procedures and continuous red-teaming exercises	Somani2025
“Defense-in-depth” across the AI Lifecycle	Tse2025
Achieve AI efficiency improvements to reduce investment costs	Wirtz2022

Note. The 24 mitigation records that could not be classified against the mitigation domain taxonomy, from 16 source documents. Includes additional domains, broad domains, and cross-cutting domains across all loci.

Supplementary Table E. Mitigation records that could not be classified to a mitigation instrument

Mitigation record	Source
Measure 8.3 Promotion of a healthy risk culture	AIOffice2025
Ensure user awareness and training	AISC2024
Characteristics of trustworthy AI are integrated into organizational policies, processes, procedures, and practices	Barrett2025
Govern 2.2: The organization's personnel and partners receive AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements	Barrett2025
Critical media literacy (1380) and education to inform the public of the capabilities and pitfalls of AI-generated content	Bengio2026
Defence in-depth	Bengio2026
Incentives for implementing cybersecurity measures, and incident reporting to qualified bodies for continuous feedback loops	Bengio2026
Insurance to restructure risk allocation and incentivise long-term safety investments (1393, 1394)	Bengio2026
Risk tolerance	Bengio2026
Societal education programmes to increase public awareness of risks and impacts	Bengio2026
Remedial interventions	Bernardi2025
Building in redundancy (for deployment mitigations)	Buhl2025
Building in redundancy (for security mitigations)	Buhl2025
Implementing comprehensive mitigations:	Buhl2025
Collaborative Governance	Crichton2025
Competence and Expertise	Crichton2025
Culture (leadership commitment)	Crichton2025
Culture (safety, cybersecurity, privacy, and risk-aware culture)	Crichton2025
Equity, Inclusion, and Access	Crichton2025
Global Governance	Crichton2025
Human-Centric Development	Crichton2025
Intellectual Property	Crichton2025
Management System	Crichton2025
Mission, Strategy, and Alignment	Crichton2025
Public Accountability	Crichton2025
Resource Allocation (for personnel and human resources)	Crichton2025
Societal and Global Stability	Crichton2025
Supply Chain Security	Crichton2025
Sustainability and Environment	Crichton2025
Training and Awareness	Crichton2025
User Awareness and Communication	Crichton2025
Workforce and Economy	Crichton2025
Implement processes and technical requirements to enable rapid rollback or withdrawal of models in case of egregious, widespread or consistent harms.	DSIT2023
In particular, recognise that even the possession of some models may be dangerous	DSIT2023
Maintain an awareness of security threats and failure modes, in particular data scientists and developers	DSIT2023
Train developers, system owners and senior leaders in secure AI practices	DSIT2023
Incentivize AI research effort	Floridi2018
Incentivize cross-disciplinary cooperation	Floridi2018
Incentivize development and use of preferable AI	Floridi2018
Incentivize ethical, legal, social considerations in research	Floridi2018
Incentivize research about public opinion	Floridi2018
Support corporate board of directors to take responsibility for AI ethics	Floridi2018
Support education and public awareness about AI	Floridi2018
6.2 Supporting Defensive Systems and Research	FMF2025
Encouraging downstream provider to evaluate models for deployment-specific failure modes	Gipiskis2024
Release strategy disagreement between developers	Gipiskis2024
AI for Enhanced Efficiency and Capability	Glasser2025
Compute Controls	Glasser2025
Economic Incentives	Glasser2025
Epistemic Security Measures	Glasser2025
Government Capacity Building	Glasser2025

Mitigation record	Source
Other Emergency Preparedness Measures	Glasser2025
Public Education	Glasser2025
Securing Critical Infrastructure	Glasser2025
User Policies and Education	Google2024
User Transparency and Controls	Google2024
AI literacy	Google2025
Enabling the ecosystem with research funding	Google2025
Defense Acceleration	Hodes2025
Increasing Fairness and Representation	Hodes2025
Multi-Stakeholder Approach	Ji2023
Open-Source Governance	Ji2023
Bug bounties	Jones2024
Regulator design	Jones2024
Reporting body	Jones2024
Reporting channels	Jones2024
Securing model parameters and other key intellectual property	Jones2024
Securing other systems	Jones2024
Societal adaptation	Jones2024
data governance	Joshi2025
Robust compliance controls	Joshi2025
strategic alignment	Joshi2025
Ethics training	Lu2024
Leadership commitment for RAI	Lu2024
A mix of features of other actors, from setting policies to adjudicating disputes to enforcing decisions	Maas2023
Advise policymakers about the existence and meaning of international law and which obligations these impose	Maas2023
Advocacy within the legislative AI policymaking process	Maas2023
AI R&D funding strategies	Maas2023
Bypass or challenge existing institutions by engaging in “competitive regime creation,” “forum shopping,” or the strategic creation of treaty conflicts	Maas2023
Campaigns or resources to educate researchers about AI risk, making AI safety research more concrete and legible, and/or creating common knowledge about researchers’ perceptions of and attitudes towards these risks	Maas2023
Compute governance	Maas2023
Coordinating behavior; reducing uncertainty, improving trust	Maas2023
Creating alternatives	Maas2023
Differential technological development (innovation prizes)	Maas2023
Employee activism and pressure, and documented communications of risks by employees (which make companies more risk averse because they are more likely to be held liable in court)	Maas2023
Framing and stigmatization around decisions or practices	Maas2023
Industry-level levers	Maas2023
Influence domestic policy processes in order to indirectly shape transnational legal processes	Maas2023
Institutional, software, and hardware mechanisms for enabling developers to make verifiable claims	Maas2023
Judge seminars and training courses	Maas2023
Lab-level (external) levers	Maas2023
Lab-level (internal) levers	Maas2023
New monitoring institutions and information markets	Maas2023
Norm entrepreneurship	Maas2023
Other inter-lab cooperation mechanisms	Maas2023
Provide services in terms of training, insurance, procurement, identification, archiving, etc.	Maas2023
Public-focused levers	Maas2023
Pursue specific career paths with key actors to contribute to good policymaking	Maas2023
Quadratic voting and radical markets	Maas2023
Sanctioning noncompliance	Maas2023
Serve as high-bandwidth policy advisor	Maas2023
Set activities and issue agendas in global AI governance institutions	Maas2023
Shape, provide, or spread narratives, ideas, “memes,” framings, or (legal) analogies for AI governance	Maas2023

Mitigation record	Source
Shaping AI research community decisions around AI governance	Maas2023
State-focused levers	Maas2023
Strengthen or reorganize existing international institutions rather than establishing new institutions	Maas2023
Various tools of "differential technology development"	Maas2023
Work to shape analogies or metaphors used by the public, policymakers, or courts in thinking about (advanced) AI	Maas2023
Work with "gatekeeper" organizations to put issues on the advocacy agenda	Maas2023
Adapt existing training programs to include modules on digital content transparency	NIST2024
Conduct joint educational activities and events in collaboration with third parties to promote best practices for managing GAI risks.	NIST2024
Develop certification programs that test proficiency in managing GAI risks and interpreting content provenance, relevant to specific industry and context.	NIST2024
Establish policies and procedures that address continual improvement processes for GAI risk measurement. Address general risks associated with a lack of explainability and transparency in GAI systems by using ample documentation and techniques such as: application of gradient-based attributions, occlusion/term reduction, counterfactual prompts and prompt engineering, and analysis of embeddings; Assess and update risk measurement approaches at regular cadences.	NIST2024
Provide input for training materials about the capabilities and limitations of GAI systems related to digital content transparency for AI Actors, other professionals, and the public about the societal impacts of AI and the role of diverse and inclusive content generation.	NIST2024
Develop a traceability management system for AI services	NTC2602024
Developers should uphold a people-centered approach, adhere to the principle of AI for good, and follow science and technology ethics	NTC2602024
Enhance the training of AI safety talents	NTC2602024
Establish and improve the mechanisms for AI safety education, industry self-regulation, and social supervision	NTC2602024
Identify and regulate outputs	NTC2602024
Prevent abuse of AI systems	NTC2602024
Promote international exchange and cooperation on AI safety governance	NTC2602024
Share information, and emergency response of AI safety risks and threats	NTC2602024
Strengthen AI supply chain security	NTC2602024
Strengthen capacity of constructing, managing, and operating AI	NTC2602024
Supply chain security	NTC2602024
Users should be aware of the potential impact of AI products on minors and take steps to prevent addiction and excessive use.	NTC2602024
Users should be mindful of cybersecurity risks when using AI products to prevent them from becoming targets of cyberattacks.	NTC2602024
Users should fully understand its data processing and privacy protection measures before using an AI product.	NTC2602024
Users should raise their awareness of the potential safety risks associated with AI products, and select AI products from reputable providers.	NTC2602024
Educational and Public Awareness Initiatives	Pathan2025
Encourage Ethical AI Research and Development	Pathan2025
Ethical Training	Pathan2025
International Collaboration	Pathan2025
AI cryptography and National Security Risks	Radanliev2025
AI firms cooperate for the public benefit	Radanliev2025
Investing in AI cybersecurity capacity as a global priority	Radanliev2025
Measures for fairness	Radanliev2025
Protection against espionage	Schuett2023
Tracking model weights	Schuett2023
Education and Public Awareness	Shi2024
Ethics and Compliance Proposals	Shi2024
Interdisciplinary Collaboration and Ethical AI Governance	Shi2024
Toward Multivalent International Cooperation and Interdisciplinary Community Building	Shi2024
Agile and adaptive governance processes and coordination	Sigfrids2022
Comprehensive frameworks on governance	Sigfrids2022
Ethics and human rights in policy making	Sigfrids2022
Long-term governance strategies	Sigfrids2022
Operationalization mechanisms	Sigfrids2022

Mitigation record	Source
Processes for improving public governance and coordination of AI	Sigfrids2022
Conduct regular training and scenario drills	Somani2025
Enforce model access, use and environmental controls	Somani2025
Enhance hardware and supply chain oversight	Somani2025
Education programs for end-users	Tkeshelashvili2025
Staff training	Tkeshelashvili2025
“Defense-in-depth” across the AI Lifecycle	Tse2025
Framework iteration cycle	Tse2025
Organizational safety culture and training	Tse2025
Policy feedback mechanism	Tse2025
Hypothetical Scenarios and Moral Dilemmas	Wang2025
Achieve AI efficiency improvements to reduce investment costs	Wirtz2022
Allow social norms to evolve steadily without hard regulation	Wirtz2022
Consider social AI impacts and ensure social equality and well-being by enhancing social safety nets (e.g., basic income), consider the entire socio-technical AI infrastructure to ensure social compliance with public interest values	Wirtz2022
Develop mechanisms to identify, prevent, and eliminate disinformation or computational propaganda	Wirtz2022
Ensure freedom of speech without algorithm-based censorship	Wirtz2022
Ensure global competitiveness and fairness and particularly support SMEs to cope with high AI costs in order to stay competitive	Wirtz2022
Ensure that AI systems support human work as a complementary tool, increase AI competencies through training, and ensure reintegration of individuals affected by technological unemployment	Wirtz2022
Ensure that individuals’ autonomy and freedom is not compromised by AI systems (e.g., decision making)	Wirtz2022
Ensure transparent communication, define goals and an AI change management strategy, support interdisciplinarity and develop an AI culture, point out AI benefits and ensure transparent handling of AI to generate higher acceptance among employees and consumers, integrate key stakeholders into implementation and change processes	Wirtz2022
Foster creation and transfer of AI knowledge (e.g., training of AI experts and education of society) to enable and realise innovation potential	Wirtz2022
Increase social ability regarding AI to support acceptance (e.g., comprehensive, interdisciplinary AI education, preparation for challenges)	Wirtz2022
Raise awareness of the ubiquity of AI and the constant collection and use of data	Wirtz2022

Note. The 171 mitigation records that could not be classified against the mitigation instrument taxonomy, from 32 source documents. Includes additional instruments, broad instruments, and cross-cutting instruments.