MIT AI Risk Index

FutureTech
THE ECONOMIC AND TECHNICAL FOUNDATIONS
OF PROGRESS IN COMPUTING

Massachusetts
Institute of
Technology

Mapping AI Risk Mitigations

Evidence Scan and Draft Mitigation Taxonomy

July 2025

Alexander K. Saeri, Sophia Lloyd George, Jess Graham,
Clelia Lacarriere, Peter Slattery

Executive Summary

What we did

- We sought to understand & organize mitigations that organizations could implement to help address risks from AI.
- We identified and extracted mitigations from documents that proposed AI risk mitigations into an **AI Risk Mitigation Database** (view on Airtable).
- We used the mitigations to develop a draft **AI Risk Mitigation Taxonomy** (see Figure 1 for overview, Appendix A for detail).

What we found

- Our evidence scan found 13 documents proposing organizational mitigations for AI risks (Included Frameworks)
- We extracted 831 mitigations from the 13 documents
- Our draft AI Risk Mitigation Taxonomy has 4 categories:
    - **Governance & Oversight Controls**: Formal organizational structures and policy frameworks that establish human oversight mechanisms and decision protocols.
    - **Technical & Security Controls**: Technical, physical, and engineering safeguards that secure AI systems and constrain model behaviors.
    - **Operational Process Controls**: Processes and management frameworks governing AI system deployment, usage, monitoring, incident handling, and validation.
    - **Transparency & Accountability Controls**: Formal disclosure practices and verification mechanisms that communicate AI system information and enable external scrutiny.
- The most common category of mitigations was Operational Process Controls (n = 295), which was mentioned in all 13 included documents.
- The draft Taxonomy has 23 subcategories. Most commonly mentioned in the included documents were Testing & Auditing (n = 127), Risk Management (n = 125), Data Governance (n = 57), Post-Deployment Monitoring (n = 50), and Risk Disclosure (n = 44).

What's next

- We will build on this evidence scan with a systematic review of AI risk mitigation frameworks (using peer-reviewed literature, gray literature, and expert consultation) and intend to revise the taxonomy based on this work.
- We welcome feedback on this evidence scan and our draft taxonomy.
- We welcome recommendations for additional frameworks or documents to include in our review.

Give feedback on the taxonomy or suggest documents to include

Figure 1. Draft AI Risk Mitigation Taxonomy

Mitigation Category	Mitigation Subcategory
1. Governance & Oversight Controls *Formal organizational structures and policy frameworks that establish human oversight mechanisms and decision protocols to ensure human accountability, ethical conduct, and risk management throughout AI development and deployment.*	1.1 Board Structure & Oversight
	1.2 Risk Management
	1.3 Conflict of Interest Protections
	1.4 Whistleblower Reporting & Protection
	1.5 Safety Decision Frameworks
	1.6 Environmental Impact Management
	1.7 Societal Impact Assessment
2. Technical & Security Controls *Technical, physical, and engineering safeguards that secure AI systems and constrain model behaviors to ensure security, safety, alignment with human values, and content integrity.*	2.1 Model & Infrastructure Security
	2.2 Model Alignment
	2.3 Model Safety Engineering
	2.4 Content Safety Controls
3. Operational Process Controls *Processes and management frameworks governing AI system deployment, usage, monitoring, incident handling, and validation, which promote safety, security, and accountability throughout the system lifecycle.*	3.1 Testing & Auditing
	3.2 Data Governance
	3.3 Access Management
	3.4 Staged Deployment
	3.5 Post-Deployment Monitoring
	3.6 Incident Response & Recovery
4. Transparency & Accountability Controls *Formal disclosure practices and verification mechanisms that communicate AI system information and enable external scrutiny to build trust, facilitate oversight, and ensure accountability to users, regulators, and the public.*	4.1 System Documentation
	4.2 Risk Disclosure
	4.3 Incident Reporting
	4.4 Governance Disclosure
	4.5 Third-Party System Access
	4.6 User Rights & Recourse

Explore an interactive version of the draft taxonomy

Refer to Appendix A for full draft taxonomy including subcategory descriptions & examples

Contents

Executive Summary **2**

What we did 2

What we found 2

What's next 2

Contents **4**

Research Motivation **5**

Methodology **5**

How we found relevant documents 5

How we identified & extracted mitigations 6

How we constructed the draft taxonomy 6

How we used AI to assist in extraction and classification 7

Insights **10**

Risk Management 10

AI risk management is an emerging concept 10

Limitations of our draft taxonomy in classifying risk management 11

Testing and Auditing 12

Categories of mitigations that were rarely mentioned 13

Future directions for research **13**

Map AI risk mitigations against the AI risks they are intended to address 14

Identify mitigations for more actors across the AI ecosystem 15

Identify organizational conditions that reduce AI risks 15

Included Frameworks in this Evidence Scan **15**

Appendix A: Draft AI Risk Mitigation Taxonomy **20**

Appendix B: Definitions **23**

Risk 23

Mitigation 23

Explicit definitions of 'risk management' from included documents 23

Research Motivation

We conducted this evidence scan to identify & synthesize emerging practice in AI risk mitigations, as part of the broader MIT AI Risk Initiative (airisk.mit.edu).

Risks from AI and proposed mitigations to reduce the likelihood or severity of these risks have been documented in policy, technical, and risk management reports. Some of this work has suggested frameworks or taxonomies to organize mitigations (e.g., NIST AI 600-1 or Eisenberg's Unified Control Framework). However, each document uses its own jargon and has gaps in coverage; it can also be difficult to know which framework fits the needs of a specific actor or decision-maker.

To address this, we identified relevant documents and extracted specific AI risk mitigations into an **AI Risk Mitigation Database**, then constructed a draft **AI Risk Mitigation Taxonomy**.

We release the database and draft taxonomy for comment & feedback now, because:

- We have observed growing demand for a comprehensive compilation of AI risk mitigations from diverse sources.
- We think it is useful to present existing mitigations in an accessible, structured way for both technical practitioners and policy stakeholders.

We intend to follow up this initial evidence scan with a systematic review of AI risk mitigation frameworks, aiming to improve coverage of the database and comprehensiveness and clarity of the mitigation taxonomy.

Overall, our intention with this work - as part of the broader MIT AI Risk Initiative - is to help individuals and organizations understand AI risks, identify and implement effective risk mitigations, and coordinate to reduce systemic and catastrophic risks.

Methodology

How we found relevant documents

Our review identified 13 relevant documents published between 2023-2025. We started with documents we had identified through previous work (e.g., an evidence scan of AI risk management frameworks; a systematic review of AI risk frameworks), then expanded using:

- Reference mining from identified frameworks
- Social media monitoring of relevant authors and organizations
- Publications from organizations focused on AI risk management

Potentially relevant documents were screened by the author team before inclusion. We included documents that proposed a framework or other structured list of mitigations for AI risks.

How we identified & extracted mitigations

We manually extracted 831 mitigations from the 13 included documents, using the following coding frame:

- Unique mitigation identifier
- Name of the mitigation (verbatim if given in document, otherwise constructed by author)
- Description of the mitigation

We defined an AI risk mitigation as *"an action that reduces the likelihood or impact of a risk"* (adapted from Society for Risk Analysis, 2018; Actuarial Standards Board, 2012; NIST SP 800-30 Rev. 1 from CNSSI 4009; see Appendix B).

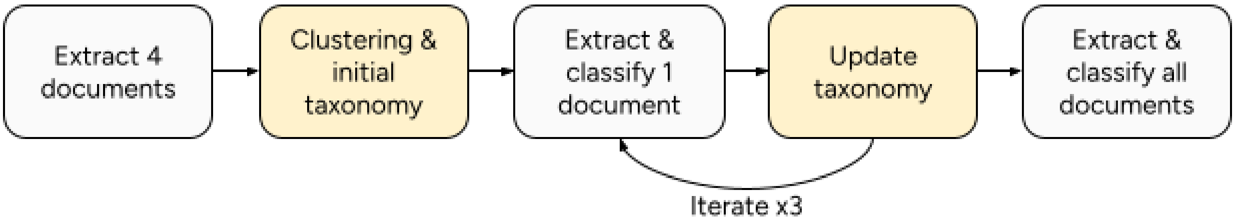
How we constructed the draft taxonomy

Overall, we had three objectives when constructing a mitigation taxonomy:

- **Practical**: We tried to cluster together mitigations that would be implemented by similar actors, took place at a similar lifecycle stage, or that involved similar implementation processes
- **Accessible**: We tried to define mitigation categories, names, and descriptions that were understandable by technical teams, policymakers, and executives
- **Comprehensive**: We tried to maximize the breadth of mitigations that could be accommodated in the taxonomy, although this reduced overall coherence

We developed our AI risk mitigation taxonomy using an iterative approach where we extracted and classified mitigations from several documents at a time. Figure 2 describes the process.

Figure 2: Overview of document extraction and taxonomy construction process


Extract 4 documents
Clustering & initial taxonomy
Extract & classify 1 document
Update taxonomy
Extract & classify all documents
Iterate x3

We first manually extracted mitigations from 4 documents, and constructed an initial taxonomy based on a thematic analysis and clustering of those mitigations. We placed significant weight on existing categories of mitigations presented in the documents. We experimented with several approaches when we conducted the initial clustering, including:

- **Risk management**: clustering by stages of risk management (identification, assessment, mitigation, etc.)
- **AI system lifecycle**: clustering by stages of AI development, deployment, and use (design, training, testing, deployment, etc.)

- **Actor-based approach**: clustering by who designs, implements, enforces, or is affected by mitigations (e.g., government agencies rolling out AI tools, companies deploying AI in their operations, AI researchers at frontier AI labs)
- **Risk-based**: clustering by the specific risk being addressed by a mitigation (e.g., as described in the MIT AI Risk Taxonomy)
- **Technical vs. socio-technical**: clustering by whether the intervention or target of a mitigation was the technical AI model/system or human/organizational behavior

After we classified all of the extracted mitigations according to the initial taxonomy, we then identified mitigations that could not be classified and gathered internal feedback within the author team. We modified the taxonomy to accommodate the unclassified mitigations, then tested the updated taxonomy on mitigations from one additional document. We repeated this process three times in total.

Over three iterations, we settled on a combination of a system lifecycle and socio-technical approach. However, we think that the other approaches described above (e.g., risk management, etc) could also be useful taxonomies to explore.

Our final draft taxonomy clustered mitigations into four categories (Governance & Oversight, Technical & Security, Operational Process, and Transparency & Accountability) and 23 subcategories (see Appendix A for the full taxonomy, or explore an interactive taxonomy).

All remaining mitigations from the documents were then classified using the final draft taxonomy. We classified 815 mitigations (98%) of the extracted mitigations using the final draft taxonomy. 16 mitigations could not be classified.

How we used AI to assist in extraction and classification

In this evidence scan, each author involved in extraction and classification experimented with LLM / AI assistants in their work. Some examples of how we used AI assistants:

- To extract mitigations from included documents, by attaching a document and providing a structured prompt instructing the AI assistant to extract relevant mitigations (Claude 4 Sonnet, Opus; ChatGPT o3)
- To experiment with mitigation category names and descriptions, standardize description format, and for feedback on edge case classifications (Claude 4 Sonnet, Opus)
- To recommend classifications against the draft taxonomy, and provide justifications for its recommendations (Claude 3.7 Sonnet, Claude 4 Sonnet)
    - ○ *Example shortened prompt: I am working with the following taxonomy of AI risk controls {draft taxonomy in XML format}. For each mitigation {mitigation name and description}, assign the best-fit category with a confidence score and justification. If no category fits, say so. List secondary categories if applicable.*

We originally planned to heavily leverage LLM / AI assistance for extraction & classification, but experienced several problems with the quality and consistency of outputs. We discovered partway through our extraction & classification of the documents that the AI assistants:

- Sometimes missed / failed to identify some mitigations from the included documents
- Occasionally combined several mitigations that appeared in the document into one
- Occasionally confabulated / hallucinated mitigations by saying that they appeared in the documents when they did not

When we identified these issues, we conducted quality assurance on the extracted and classified mitigations through a detailed audit of the mitigations database. This involved manually reviewing the included documents to verify that all mitigations had been accurately extracted and that no spurious entries had been added.

After verifying the extractions, a member of the author team reviewed all classifications. This involved reading each mitigation name & description, reviewing the AI-proposed classification & justification, and either accepting the proposal or changing the classification (and documenting their justification for doing so). Multiple team members cross-checked classifications to ensure consistency.

We include this report on the problems of using AI assistance to accelerate evidence synthesis for transparency and to highlight the current need for careful human validation of AI assistance to ensure that research findings are accurate.

The distribution of mitigations across each of our taxonomy categories and subcategories is shown in Table 1.

Table 1. AI Mitigations Database Coded With Draft AI Risk Mitigation Taxonomy

Category / Subcategory		Number of Mitigations	Percentage of Mitigations	Number of Documents
1	**Governance & Oversight Controls**	**248**	**30%**	**13**
1.1	Board Structure & Oversight	35	4%	8
1.2	Risk Management	125	15%	13
1.3	Conflict of Interest Protections	8	1%	3
1.4	Whistleblower Reporting & Protection	10	1%	7
1.5	Safety Decision Frameworks	31	4%	11
1.6	Environmental Impact Management	11	1%	5
1.7	Societal Impact Assessment	28	3%	7
2	**Technical & Security Controls**	**101**	**12%**	**12**
2.1	Model & Infrastructure Security	32	4%	11
2.2	Model Alignment	9	1%	5
2.3	Model Safety Engineering	38	5%	7
2.4	Content Safety Controls	22	3%	6
3	**Operational Process Controls**	**295**	**36%**	**13**
3.1	Testing & Auditing	127	15%	12
3.2	Data Governance	57	7%	7
3.3	Access Management	23	3%	7
3.4	Staged Deployment	8	1%	5
3.5	Post-deployment Monitoring	50	6%	9
3.6	Incident Response & Recovery	30	4%	9
4	**Transparency & Accountability Controls**	**171**	**21%**	**13**
4.1	System Documentation	37	4%	10
4.2	Risk Disclosure	44	5%	11
4.3	Incident Reporting	30	4%	11
4.4	Governance Disclosure	24	3%	9
4.5	Third-Party System Access	19	2%	7
4.6	User Rights & Recourse	19	2%	5
X.X	Mitigation not otherwise categorized	16	2%	8
	TOTAL	831	100%	13

Insights

In this section, we discuss:

- Our observations about AI risk management based on the included documents, and limitations in our taxonomy for the 1.2 Risk Management category
- Our observations about testing & auditing (e.g., red teaming) based on the included documents, and limitations in our taxonomy for the 3.1 Testing & Auditing category
- Categories of mitigations that were rarely mentioned in the included documents

Risk Management

We classified 125 (15%) of all identified mitigations as 'risk management' using our draft taxonomy, but there were many different processes and actions that could be classified under this umbrella term.

We believe that this reflects that 'AI risk management' is an emerging concept; the boundaries of risk management - and what should be included or excluded as a risk management action - are not yet settled. In addition, we believe that the category of risk management in our draft taxonomy needs to be decomposed and clarified.

AI risk management is an emerging concept

In 2023, Schuett et al. conducted an expert survey of leaders from AGI companies, academia, and civil society. Experts were asked to agree, disagree, or express uncertainty about 50 different AGI safety practices. Of all the practices surveyed, enterprise risk management was the statement with the highest "I don't know" rate (26%); as Schuett et al. remark, this "indicates that many respondents simply did not know what enterprise risk management is and how it works".

In contrast, some of the included documents in our evidence scan explicitly defined risk management (e.g., NIST, 2024; Campos et al., 2025; Gipiškis et al., 2024; Bengio et al., 2025; see Appendix B for details).

These definitions typically included several related stages or functions, including *identifying* risks, *assessing / evaluating* risks, and *implementing measures* to reduce risks. Some also emphasized the role of *governance* - rules, procedures, or culture - to provide structure to or sustain the other stages or functions, or *monitoring* AI system behavior once it is deployed and being used.

On the basis of these descriptions, we defined **1.2 Risk Management** in our draft taxonomy as:

> *Systematic methods that identify, evaluate, and manage AI risks for comprehensive risk governance across organizations.*

However, once we started classifying mitigations from the included documents, we found that many of them could be classified as risk management, based on this definition. These included mitigations as varied as conducting adequacy assessments of a lab's safety and security frameworks (EU AI Office 2024); implementing System-Theoretic Process Analysis (STPA;

Gipiškis 2024); and updating due diligence processes when purchasing generative AI systems (NIST, 2024).

Our finding that many mitigations can be classified under risk management, the emerging conceptualization of AI risk management, and the inconsistency between current understanding (Schuett et al., 2023) and best-practice proposals (e.g., NIST, 2024; Campos et al., 2025), suggests two potential challenges to effectively addressing risks from AI:

- Organizations may implement nearly any action that identifies, evaluates, and/or mitigates AI risks and believe (or claim) that they are engaging in 'risk management'
- Conflicting frameworks and definitions of 'risk management' makes it more difficult to coordinate effective action

Limitations of our draft taxonomy in classifying risk management

Our draft taxonomy, developed through iteration to comprehensively classify all of the mitigations extracted from the included documents, likely requires further decomposition of risk management as a (sub)category of mitigations. For example, other existing subcategories have conceptual overlap with the specific functions of risk management (e.g., identifying, analyzing, enacting measures, governing and/or monitoring), including safety decision frameworks (1.5), risk disclosure (4.2), and post-deployment monitoring (3.5).

Table 2: Categories of mitigations conceptually similar to 1.2 Risk Management

Subcategory	Function	Examples
1.5 Safety Decision Frameworks	Policy-setting mechanisms that constrain AI development and deployment	*If-then safety protocols, capability ceilings, deployment pause triggers, safety-capability resource ratios*
4.2 Risk Disclosure	Communication of risk information to external stakeholders	*Publishing risk assessment summaries, pre-deployment notifications to government, reporting large training runs, disclosing mitigation strategies, notifying affected parties*
3.5 Post-deployment Monitoring	Ongoing tracking of deployed AI systems	*User interaction tracking systems, capability evolution assessments, periodic impact reports, automated misuse detection, usage pattern analysis tools*
1.7 Societal Impact Assessment	Evaluation of AI systems' broader effects on society and communities	*Fundamental rights impact assessments, expert consultations on risk domains, stakeholder engagement processes, governance gap analyzes*

We would welcome feedback to disambiguate 'risk management' as a category of mitigations. We are considering the following directions:

1. **Emphasizing risk assessment** as a distinct category focused on identification and evaluation.
2. **Partitioning off** risk treatment and mitigation implementation into separate categories.
3. **Creating clearer boundaries** between categories (i.e. safety decision frameworks, risk disclosure, governance disclosure, post-deployment monitoring, and societal impact assessment, amongst other subcategories) to distinguish between the key phases of risk management.

Testing and Auditing

We classified 127 (15%) of all identified mitigations as 'testing & auditing' using our draft taxonomy; it was the most commonly mentioned subcategory of mitigations.

We defined **3.1 Testing & Auditing** in our draft taxonomy as:

> *Systematic internal and external evaluations that assess AI systems, infrastructure, and compliance processes to identify risks, verify safety, and ensure performance meets standards.*

As with risk management, we noticed that testing & auditing could be used to classify many different actions proposed in the included documents. Extractions from the documents indicate that testing and auditing can be quite wide-ranging, encompassing red teaming, external and internal audits, benchmarks, and bug bounty programs (Campos 2025; Schuett et al. 2023).

In addition, other existing subcategories have conceptual overlap with some of the activities described under testing & auditing, including model risk management (1.2), safety engineering (2.3), and third-party system access (4.5).

Table 3: Categories of mitigations conceptually similar to 3.1 Testing & Auditing

Subcategory	Function	Examples
1.2 Risk Management	Systematic evaluation checkpoints for risk assessment	*Pre-deployment risk assessments, independent risk assessments*
2.3 Model Safety Engineering	Technical methods for safety constraints	*Safety analysis protocols, hierarchical auditing*
4.5 Third-Party System Access	Infrastructure enabling external validation	*Third-party capability assessments, researcher access programs, government access provisions*

We would welcome feedback to disambiguate 'Testing & Auditing' as a category of mitigations. We are considering the following directions:

1. **Partitioning based on lifecycle stage** (e.g., pre-deployment vs. post-deployment)
2. **Distinguishing between who conducts the evaluation** (e.g., internal vs. external)
3. **Specifying assessment focus** (technical capabilities vs. safety compliance vs. societal impact)

Categories of mitigations that were rarely mentioned

Several categories of mitigations received limited attention in the included documents. We wish to stress that the frequency of with which a mitigation is classified does not necessarily reflect its relative importance or rate of adoption in practice; we also recognize that our decisions in drawing category boundaries will influence the relative frequency of mitigations in each category.

However, these findings suggest that some of these categories of mitigation could be investigated, developed, and defined in more detail. Table 4 presents these mitigation categories.

We wish to draw special attention to *2.2 Model Alignment*. Actions to align AI models and systems with human goals & values were rarely mentioned, despite their fundamental importance in addressing catastrophic and existential risks from advanced AI systems.

Table 4: Categories of mitigations that were rarely mentioned in the included documents

Subcategory	Function	% mitigations (# documents)
1.3 Conflict of Interest Protections	Manage financial interests and organizational structures to ensure leadership can prioritize safety over profit motives in critical situations	<1% (3)
1.4 Whistleblower Reporting & Protection	Enable confidential reporting of safety concerns or ethical violations	<1% (7)
1.6 Environmental Impact Management	Measuring, reporting, and reducing the environmental footprint of AI systems	<1% (5)
2.2 Model Alignment	Ensure AI systems understand and adhere to human values and intentions	<1% (5)
3.4 Staged Deployment	Deploy AI systems in stages, requiring safety validation before expanding user access or capabilities	<1% (5)

Future directions for research

Our evidence scan, development of the draft AI risk mitigations database, and construction of the draft taxonomy, suggests several directions for applied research. We welcome feedback on these proposed directions.

Map AI risk mitigations against the AI risks they are intended to address

Many of the mitigations extracted from the documents did not specify what type of AI risk they were intended to address. This is a problem, given that an actor that seeks to address *misinformation* risks would likely need to implement a different mitigation than one that seeks to address *AI possessing dangerous capabilities*. Not all actors are vulnerable to or accountable for addressing all AI risks.

In our previous work, the AI Risk Repository, we constructed taxonomies to classify risks: by causal factor and by domain of harm. These and other taxonomies of risk are starting to be mapped to other structured knowledge about AI, including legislation and standards (e.g., ETO AGORA), incidents of harm (e.g., AI Incident Database), and comprehensive ontologies to structure knowledge on AI risks (e.g., the AI Risk Ontology; The IBM Risk Atlas Nexus).

We strongly believe that a **mapping of AI risks to mitigations is needed to advance coordinated action on AI risks**. Some of the included documents do propose relationships between risks and mitigations, including NIST AI 600-1 (2024) and especially Gipiškis et al. (2024). Gipiškis and colleagues identify risk sources and link them to corresponding risk measures; a promising approach for addressing this gap. Developing or linking the risk sources and measures to taxonomies and ontologies can help to increase its interoperability with other work that seeks to structure knowledge on AI risks and mitigations. Table 5 illustrates examples of risk sources and measures from Gipiškis et al. (2024).

Table 5: Examples of Risk Sources & Measures (Gipiškis et al. 2024)

Risk Source (Gipiškis et al. 2024)	Risk Measure (Gipiškis et al. 2024)
Difficulty filtering large web scrapes or large scale web datasets.	Use of synthetic data
Benchmark leakage or data contamination	Reporting Data Decontamination efforts Contamination Detection
Adversarial Examples	Adversarial Training
Fine-tuning dataset poisoning	Data cleaning; internal data poisoning diagnosis
Inaccurate measurement of model encoded human values; Biased evaluations of encoded human values	Frequent Benchmarking to identify when red teaming is needed; Frequent testing when scaling model or dataset

Note: *this table represents only a subset of all the risk sources and corresponding measures identified by Gipiškis et al. 2024, sampled for illustrative purposes.*

Identify mitigations for more actors across the AI ecosystem

In this evidence scan, we searched for documents that discussed AI risk mitigations. The documents we found and included tended to focus on organizations that were developing and deploying the most advanced AI systems (e.g., “frontier AI”, “generative AI”, “general-purpose AI”). However, we lack an understanding of actions that other actors could take to mitigate AI risks.

It could be useful to identify proposed and enacted actions by these actors, and how they reinforce, substitute, or undermine others’ actions to mitigate AI risks.

A non-exhaustive set of examples of such actors whose actions could be investigated include:

- Purchasers and users of advanced AI, such as large corporate organizations
- Regulators of advanced AI, such as government agencies
- People and groups subject to the use of AI by others, such as customers of an online retailer who must converse with an AI assistant to return a faulty item

Identify organizational conditions that reduce AI risks

In risk management frameworks and in our draft taxonomy, there is a focus on measurable and observable actions that fit within a typical conceptualization of what organizations ‘do’.

We think that investigating the organizational conditions that reduce risks from AI could be an important contribution and one that sits in tension with our focus on classifying mitigations, which naturally privileges observable actions.

One example of supportive organizational conditions is a positive *company safety culture*, mentioned in several included documents but especially in Campos (2025) where it is framed as part of the *risk governance* component of frontier AI risk management.

Other examples could include:

- Defusing competitive pressure (within or between organizations)
- Supporting psychological safety (e.g., feeling secure and perceiving positive norms to ‘speak up’, promote safe practice, or criticize unsafe practice)
- How teams within organizations are structured or managed
- How effectively safety practices are implemented within an organization.

Included Frameworks in this Evidence Scan

1. International AI Safety Report (Bengio et al., 2025)
2. A Frontier AI Risk Management Framework: Bridging the gap between current AI practices and established risk management (Campos et al., 2025)
3. The Unified Control Framework (Eisenberg et al., 2025)
4. Risk Sources and Risk Management Measures in support of standards for general-purpose AI systems (Gipiškis et al., 2024)
5. Effective Mitigations for Systemic Risks from General-Purpose AI (Uuk et al., 2024)
6. FLI AI Safety Index 2024 (Future of Life Institute, 2024)
7. Towards Best Practices in AGI Safety and Governance (Schuett et al., 2023)
8. Pitfalls of Evidence-Based AI Policy (Casper et al., 2025)
9. EU AI Act: General Purpose AI Code of Practice (Draft 3) (EU AI Office, 2025)
10. Emerging Processes for Frontier AI Safety (UK Government, 2023)
11. NIST AI Risk Management Framework: Generative AI Profile (NIST, 2024)
12. AI Risk Management Standards Profile for GPAIS (Barrett et al., 2024)
13. California Senate Bill 1047 (Wiener, 2024)

Access detailed bibliographic information about included documents on Airtable

Access all included documents in a public Paperpile folder

International AI Safety Report (Bengio et al., 2025)

The first comprehensive international synthesis of evidence on AI risks and their relevant technical mitigations, mandated by nations attending the AI Safety Summit in Bletchley Park. This report represents a global collaboration of 30 nations and is the culmination of 100 AI experts' efforts to establish a common understanding of frontier AI systems.

Bengio, Y., et al. (2025). *International AI safety report.* UK AI Safety Institute. https://www.gov.uk/government/publications/international-ai-safety-report-2025

A Frontier AI Risk Management Framework: Bridging the gap between current AI practices and established risk management (Campos et al., 2025)

A recent framework proposing an integrated risk management strategy for frontier AI risks, emphasizing the gap between current AI practices and established risk management methodologies from high-risk industries. The approach consists of four stages: (1) risk identification, (2) risk analysis and evaluation, (3) risk treatment, and (4) risk governance.

Campos, M., Stewart, A., & Zhang, R. (2025). *A frontier AI risk management framework: Bridging the gap between current AI practices and established risk management.* arXiv. https://arxiv.org/pdf/2502.06656

The Unified Control Framework (Eisenberg et al., 2025)

A comprehensive framework establishing 42 controls for enterprise AI governance, risk management, and regulatory compliance. The framework unifies fragmented approaches into a cohesive system for managing AI risks across organizations.

Eisenberg, C., Seaver, A., & Rubin, J. (2025). The unified control framework: Establishing a common foundation for enterprise AI governance, risk management and regulatory compliance. arXiv. https://arxiv.org/abs/2503.05937

Risk Sources and Risk Management Measures in support of standards for general-purpose AI systems (Gipiškis et al., 2024)

A systematic analysis identifying key risk sources for general-purpose AI systems and corresponding management measures, developed to support global standardization efforts.

Gipiškis, D., et al. (2024). *Risk sources and risk management measures in support of standards for general-purpose AI systems.* arXiv. https://arxiv.org/abs/2410.23472

Effective Mitigations for Systemic Risks from General-Purpose AI (Uuk et al., 2024)

A survey of experts ranging from AI safety to CBRN risks to determine the most effective risk management practices for general-purpose AI models. Three mitigations emerge as especially salient: safety incident reports and security information disclosure, third-party pre-deployment model audits, and pre-deployment risk assessments. Policy guidance is offered for delivering the most crucial measures to mitigate risks from frontier AI systems.

Uuk, R., Lam, H., Simeon, V., & O'Brien, N. (2024). *Effective mitigations for systemic risks from general-purpose AI.* arXiv. https://doi.org/10.48550/arXiv.2412.02145

FLI AI Safety Index 2024 (Future of Life Institute, 2024)

The 2024 FLI AI Safety Index conducted a comprehensive review of six prominent AI companies by an independent panel of seven technical AI and governance specialists. Key findings include significant risk management discrepancies across the companies, vulnerability to adversarial attacks, lack of control protocols, and insufficient external oversight. Mitigation strategies are recommended to counteract these AI risk deficiencies.

Future of Life Institute. (2024). *FLI AI safety index 2024*. Future of Life Institute. https://futureoflife.org/document/fli-ai-safety-index-2024/

Towards Best Practices in AGI Safety and Governance (Schuett et al., 2023)

A survey of expert opinion from leading AI companies including OpenAI, Google DeepMind, and Anthropic on best practices for AGI safety and governance. The paper identifies 51 specific practices ranging from technical safety measures to governance structures.

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. https://arxiv.org/abs/2305.07153

Pitfalls of Evidence-Based AI Policy (Casper et al., 2025)

A critique of the outsized emphasis on evidence-based AI policy, which can discourage regulatory action in the face of uncertain risks. Challenging such stagnancy, the authors maintain that uncertainty is cause for passing regulation in the near term. They recommend mitigation strategies that are grounded in but not overreliant on empirical findings, spanning AI governance institutes to shutdown procedures for AI systems.

Casper, S., et al. (2025). *Pitfalls of evidence-based AI policy.* arXiv. https://arxiv.org/abs/2502.09618

EU AI Act: General Purpose AI Code of Practice (Draft 3) (EU AI Office, 2025)

The implementing guidance for the EU AI Act's requirements on general-purpose AI systems, detailing specific measures for systemic risk assessment, mitigation, and governance.

EU AI Office. (2025). *EU AI Act: General purpose AI code of practice (Draft 3).* European Commission. https://code-of-practice.ai/

Emerging Processes for Frontier AI Safety (UK Government, 2023)

A comprehensive catalog of safety practices being implemented or considered by frontier AI organizations, providing transparency into current industry approaches to risk management.

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

NIST AI Risk Management Framework: Generative AI Profile (NIST, 2024)

US National Institute of Standards and Technology (NIST) guidance adapting the AI Risk Management Framework specifically for generative AI, containing detailed controls and implementation guidelines.

National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1)*. NIST. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

AI Risk Management Standards Profile for GPAIS (Barrett et al., 2024)

A profile aligned with the US NIST AI Risk Management Framework providing risk management standards specifically tailored for general-purpose AI systems and foundation models, bridging multiple existing frameworks.

Barrett, A. M., Newman, J., Nonnecke, B., Hendrycks, D., Murphy, E. R., & Jackson, K. (2024). *AI risk management standards profile for GPAIS*. UC Berkeley Center for Long-Term Cybersecurity. https://cltc.berkeley.edu/wp-content/uploads/2023/11/Berkeley-GPAIS-Foundation-Model-Risk-Management-Standards-Profile-v1.0.pdf

California Senate Bill 1047 (Wiener, 2024)

Proposed legislation establishing safety requirements for organizations developing frontier AI models - specifically, models which cost more than $100 million to train, with compute of >10^26 FLOP (or fine-tuned models of similar size). Though ultimately vetoed by the Governor of California, SB 1047 set important precedents for AI regulation.

Wiener, S. (2024). California State Senate Bill 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act. California State Legislature. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB1047

Appendix A: Draft AI Risk Mitigation Taxonomy

Mitigation Category	Mitigation Subcategory	Subcategory description	Examples
1. Governance & Oversight Controls *Formal organizational structures and policy frameworks that establish human oversight mechanisms and decision protocols to ensure human accountability, ethical conduct, and risk management throughout AI development and deployment.*	1.1 Board Structure & Oversight	Governance structures and leadership roles that establish executive accountability for AI safety and risk management.	*Dedicated risk committees, safety teams, ethics boards, crisis simulation training, multi-party authorization protocols, deployment veto powers*
	1.2 Risk Management	Systematic methods that identify, evaluate, and manage AI risks for comprehensive risk governance across organizations.	*Enterprise risk management frameworks, risk registers with capability thresholds, compliance programs, pre-deployment risk assessments, independent risk assessments*
	1.3 Conflict of Interest Protections	Governance mechanisms that manage financial interests and organizational structures to ensure leadership can prioritize safety over profit motives in critical situations.	*Background checks for key personnel, windfall profit redistribution plans, stake limitation policies, protections against shareholder pressure*
	1.4 Whistleblower Reporting & Protection	Policies and systems that enable confidential reporting of safety concerns or ethical violations to prevent retaliation and encourage disclosure of risks.	*Anonymous reporting channels, non-retaliation guarantees, limitations on non-disparagement agreements, external whistleblower handling services*
	1.5 Safety Decision Frameworks	Protocols and commitments that constrain decision-making about model development, deployment, and capability scaling, and govern safety-capability resource allocation to prevent unsafe AI advancement.	*If-then safety protocols, capability ceilings, deployment pause triggers, safety-capability resource ratios*
	1.6 Environmental Impact Management	Processes for measuring, reporting, and reducing the environmental footprint of AI systems to ensure sustainability and responsible resource use.	*Carbon footprint assessment, emission offset programs, energy efficiency optimization, resource consumption tracking*
	1.7 Societal Impact Assessment	Processes that assess AI systems' effects on society, including impacts on employment, power dynamics, political processes, and cultural values.	*Fundamental rights impact assessments, expert consultations on risk domains, stakeholder engagement processes, governance gap analyses*

Mitigation Category	Mitigation Subcategory	Subcategory description	Examples
2. Technical & Security Controls *Technical, physical, and engineering safeguards that secure AI systems and constrain model behaviors to ensure security, safety, alignment with human values, and content integrity.*	2.1 Model & Infrastructure Security	Technical and physical safeguards that secure AI models, weights, and infrastructure to prevent unauthorized access, theft, tampering, and espionage.	*Model weight tracking systems, multi-factor authentication protocols, physical access controls, background security checks, compliance with information security standards*
	2.2 Model Alignment	Technical methods to ensure AI systems understand and adhere to human values and intentions.	*Reinforcement learning from human feedback (RLHF), direct preference optimization (DPO), constitutional AI training, value alignment verification systems*
	2.3 Model Safety Engineering	Technical methods and safeguards that constrain model behaviors and protect against exploitation and vulnerabilities.	*Safety analysis protocols, capability restriction mechanisms, hazardous knowledge unlearning techniques, input/output filtering systems, defense-in-depth implementations, adversarial robustness training, hierarchical auditing, action replacement*
	2.4 Content Safety Controls	Technical systems and processes that detect, filter, and label AI-generated content to identify misuse and enable content provenance tracking.	*Synthetic media watermarking, content filtering mechanisms, prohibited content detection, metadata tagging protocols, deepfake creation restrictions*
3. Operational Process Controls *Processes and management frameworks governing AI system deployment, usage, monitoring, incident handling, and validation, which promote safety, security, and accountability throughout the system lifecycle.*	3.1 Testing & Auditing	Systematic internal and external evaluations that assess AI systems, infrastructure, and compliance processes to identify risks, verify safety, and ensure performance meets standards.	*Third-party audits, red teaming, penetration testing, dangerous capability evaluations, bug bounty programs*
	3.2 Data Governance	Policies and procedures that govern responsible data acquisition, curation, and usage to ensure compliance, quality, user privacy, and removal of harmful content.	*Harmful content filtering protocols, compliance checks for data collection standards, user data privacy controls, data curation processes*
	3.3 Access Management	Operational policies and verification systems that govern who can use AI systems and for what purposes to prevent safety circumvention, deliberate misuse, and deployment in high-risk contexts.	*KYC verification requirements, API-only access controls, fine-tuning restrictions, acceptable use policies, high-stakes application prohibitions*
	3.4 Staged Deployment	Implementation protocols that deploy AI systems in stages, requiring safety validation before expanding user access or capabilities.	*Limited API access programs, gradual user base expansion, capability threshold assessments, pre-deployment validation checkpoints, treating model updates as new deployments*
	3.5 Post-Deployment Monitoring	Ongoing monitoring processes that track AI behavior, user interactions, and societal impacts post-deployment to detect misuse, emergent dangerous capabilities, and harmful effects.	*User interaction tracking systems, capability evolution assessments, periodic impact reports, automated misuse detection, usage pattern analysis tools*

Mitigation Category	Mitigation Subcategory	Subcategory description	Examples
	3.6 Incident Response & Recovery	Protocols and technical systems that respond to security incidents, safety failures, or capability misuse to contain harm and restore safe operations.	*Incident response plans, emergency shutdown/rollback procedures, model containment mechanisms, safety drills, critical infrastructure protection measures*
4. Transparency & Accountability Controls *Formal disclosure practices and verification mechanisms that communicate AI system information and enable external scrutiny to build trust, facilitate oversight, and ensure accountability to users, regulators, and the public.*	4.1 System Documentation	Comprehensive documentation protocols that record technical specifications, intended uses, capabilities, and limitations of AI systems to enable informed evaluation and governance.	*Model cards, system architecture documentation, compute resource disclosures, safety test result reports, system prompts, model specifications*
	4.2 Risk Disclosure	Formal reporting protocols and notification systems that communicate risk information, mitigation plans, safety evaluations, and significant AI activities to enable external oversight and inform stakeholders.	*Publishing risk assessment summaries, pre-deployment notifications to government, reporting large training runs, disclosing mitigation strategies, notifying affected parties*
	4.3 Incident Reporting	Formal processes and protocols that document and share AI safety incidents, security breaches, near-misses, and relevant threat intelligence with appropriate stakeholders to enable coordinated responses and systemic improvements.	*Cyber threat intelligence sharing networks, mandatory breach notification procedures, incident database contributions, cross-industry safety reporting mechanisms, standardized near-miss documentation protocols*
	4.4 Governance Disclosure	Formal disclosure mechanisms that communicate governance structures, decision frameworks, and safety commitments to enhance transparency and enable external oversight of high-stakes AI decisions.	*Published safety and/or alignment strategies, governance documentation, safety cases, model registration protocols, public commitment disclosures*
	4.5 Third-Party System Access	Mechanisms granting controlled system access to vetted external parties to enable independent assessment, validation, and safety research of AI models and capabilities.	*Researcher access programs, third-party capability assessments, government access provisions, legal safe harbors for public interest evaluations*
	4.6 User Rights & Recourse	Frameworks and procedures that enable users to identify and understand AI system interactions, report issues, request explanations, and seek recourse or remediation when affected by AI systems.	*User reporting channels, appeal processes, explanation request systems, remediation protocols, content verification*

Appendix B: Definitions

Risk

We define a **‘risk’** as the possibility of an unfortunate occurrence

- After Society for Risk Analysis Glossary (2018)

  “Risk is the possibility of an unfortunate occurrence”

Mitigation

We define a **‘mitigation’** as an action that reduces the likelihood or impact of a risk

- After Society for Risk Analysis Glossary (2018)

  “Process of actions to reduce risk”
- And after Actuarial Standards Board standard of practice #46: Risk evaluation in enterprise risk management (2012)

  “Action that reduces the frequency or severity of a risk”
- And after NIST SP 800-30 Rev. 1 from CNSSI 4009

  “Prioritizing, evaluating, and implementing the appropriate risk-reducing controls/countermeasures recommended from the risk management process”

Explicit definitions of ‘risk management’ from included documents

Bengio et al. (2025), in *International AI Safety Report 2025*, defined AI risk management as:

> *The systematic process of identifying, evaluating, mitigating and monitoring risks.*

Campos et al. (2025), in *A Frontier AI Risk Management Framework* propose a framework for frontier AI risk management that includes four components:

> ***Risk identification*** *is the process of identifying risks and understanding their nature (i.e. risk sources and risk scenarios).*
>
> ***Risk analysis and evaluation*** *is a process that starts with the definition of a risk tolerance. This risk tolerance is then operationalized into risk indicators and their corresponding mitigations required to reduce risk below the risk tolerance.*
>
> ***Risk treatment*** *corresponds to the process of determining, implementing, and evaluating appropriate risk-reducing countermeasures.*
>
> ***Risk governance*** *corresponds to the rules and procedures that structure the risk management system in terms of decision-making, responsibilities, authority, and accountability mechanisms.*

NIST (2024), in *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)* describes AI risk management as including four functions (consistent with NIST AI 100-1):

> ***Govern****: A culture of risk management is cultivated and present*
>
> ***Map****: Context is recognized and risks related to context are identified*
>
> ***Measure****: Identified risks are assessed, analyzed, or tracked*
>
> ***Manage****: Risks are prioritized and acted upon based on a projected impact*

Gipiškis et al (2024) define an AI risk management measure (which is embedded within a broader safety engineering process) as:

> *a measure that is designed to lower risk, either when applied alone or in combination with other measures. This can include identification, mitigation, or prevention of risk sources or individual risks relevant to a given system or class of systems*