Mapping Frameworks at the Intersection of AI Safety and Traditional Risk Management

EVIDENCE SCAN

February 2025

Authors

Alexander K. Saeri, Peter Slattery and Jess Graham

Mapping Frameworks at the Intersection of AI Safety and Traditional Risk Management

Insights

- Our evidence scan found 11 AI risk management frameworks at the intersection of traditional risk management and AI safety
- All frameworks were from 2023 or newer, and are a mix of preprints, reports, government guidance documents, research and conference papers; primary authors are from UK, Singapore, Germany, Finland, USA and France.
- Terms used for AI are "frontier AI", "general-purpose AI", "advanced AI", "high-risk AI" and "Artificial General Intelligence".

Our Jan 2025 search found four main categories of frameworks at the intersection of Traditional Risk Management and AI Safety, summarised in the table:

Category	**Number**	**Core Question**	**Example**
Risk Management Translation	5	"What insights can we apply from traditional risk management for AI?"	A Frontier AI Risk Management Framework [SaferAI, 2025]
Maturity models	2	"How can we assess the maturity of organizational AI risk management?"	Framework to Rate AI Developers' Risk Management Maturity [SaferAI, 2024]
Novel approaches	3	"What new methods could address AI-specific risks?"	Affirmative safety [Wasil et al., 2024]
Emerging practice	1	"How are organizations actually managing AI risks?"	Emerging Processes for Frontier AI Safety [UK DSIT, 2023]

This 'evidence scan' aims to:

1. Identify and describe existing frameworks that combine Traditional Risk Management and AI Safety approaches
2. Help others by:
   - Making these frameworks more widely known
   - Connecting framework creators to encourage collaboration
   - Explaining current research and practices
   - Matching users with the right frameworks
   - Preventing duplicate work by consolidating existing knowledge

For access to full texts, citation details, and PDFs where available, all documents are compiled in a public Paperpile folder.

Research Motivation

Traditional risk management is a well-established discipline with robust frameworks for identifying, assessing, and mitigating risks across institutions. While traditional risk management approaches have proven effective for many contexts, they often struggle with the kinds of risks posed by AI systems - particularly those that could have very high severity, but have uncertain likelihood and reach (exposure).

This lack of integration between fields means both are missing valuable insights - traditional risk management lacks vital perspectives on AI risks and uncertainty, while AI safety work hasn't benefited from decades of proven, practical risk management methods. There's an emerging need to bridge this gap, particularly as organizations begin incorporating AI risks into their conventional risk management processes, and frontier AI labs increasingly need robust, practical frameworks to manage unprecedented risks.

Methodology

We conducted an evidence scan of frameworks that addressed **advanced AI systems** and attempted to **combine traditional risk management principles with AI safety approaches**. We used a snowball sampling approach starting from known "seed" frameworks that fit our criteria. We expanded our search by:

1. Mining reference lists of identified frameworks
2. Following content shared by relevant authors on social media platforms
3. Monitoring publications from organizations working on AI risk management and governance

This approach allowed us to identify emerging frameworks in what is a rapidly evolving field, though we acknowledge it is likely not exhaustive. Our inclusion criteria focused on frameworks that explicitly attempted to bridge the gap between traditional risk management and AI safety considerations, particularly for advanced AI systems.

Included Frameworks

Risk management translation

The frameworks in this category describe or adapt risk management methods from other safety-critical domains like cybersecurity, aviation, and nuclear power to address the risks of general-purpose or frontier AI systems. They tend to adapt governance structures (e.g., 'three lines of defense'), systematic risk assessment techniques (e.g., scenario analysis), and map risk management activities to standards, processes, and/or controls from existing frameworks (e.g., NIST, ISO).

A Frontier AI Risk Management Framework: Bridging the Gap Between Current AI Practices and Established Risk Management [SaferAI, 2025]

> This paper presents a comprehensive risk management framework for the development of frontier AI that integrates established risk management principles with emerging AI-specific practices. The framework consists of four key components: (1) risk identification (through literature review, open-ended red-teaming, and risk modeling), (2) risk analysis and evaluation using quantitative metrics and clearly defined thresholds, (3) risk treatment through mitigation measures such as containment, deployment controls, and assurance processes, and (4) risk governance establishing clear organizational structures and accountability.

Campos, S., Papadatos, H., Roger, F., Touzet, C., Murray, M., & Quarks, O. (2025). *A frontier AI risk management framework: Bridging the gap between current AI practices and established risk management*. In arXiv [cs.AI]. arXiv. http://arxiv.org/abs/2502.06656

Risk Assessment at AGI Companies: A review of popular risk assessment techniques from other safety-critical industries [GovAI, 2023]

> This paper reviews best practice risk assessment techniques from safety-critical industries and suggests ways in which organisations developing advanced AI could use them to assess catastrophic risks from AI. The paper discusses three risk identification techniques (scenario analysis, fishbone method, risk typologies and taxonomies), five risk analysis techniques (causal mapping, Delphi technique, cross-impact analysis, bow tie analysis, and system-theoretic process analysis), and two risk evaluation techniques (checklists and risk matrices).

Koessler, L., & Schuett, J. (2023). *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries*. Centre for the Governance of AI. http://arxiv.org/abs/2307.08823

AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models [Center For Long-Term Cybersecurity, 2023]

This document provides risk-management practices or controls for identifying, analyzing and mitigating risks of general-purpose AI systems. It is tailored to complement other AI risk management standards, such as the NIST AI Risk Management Framework and ISO/IEC 23894, and can provide GPAIS deployers, evaluators, and regulators with information useful for evaluating the extent to which developers of such AI systems have followed relevant best practices.

Barrett, A. M. Newman, J. Nonnecke, B. Hendrycks, D. Murphy, E. R. Jackson, K. (2023). *AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models.* UC Berkeley Center For Long-Term Cybersecurity. https://cltc.berkeley.edu/publication/ai-risk-management-standards-profile/

Transforming Risk Governance at Frontier AI Companies [Center for Long-Term Resilience, 2024]

This report explores how aspects of best practice risk governance – particularly the Three Lines Model (3LoD), which separates risk ownership, oversight and audit – could be effectively implemented at frontier AI companies to ensure safer model development and deployment.

Robinson, B., & Ginns, J. (2024). *Transforming risk governance at frontier AI companies.* The Centre for Long-Term Resilience. https://www.longtermresilience.org/wp-content/uploads/2024/07/Transforming-risk-governance-at-frontier-AI-companies-CLTR-1.pdf

Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach [Institute for AI Policy & Strategy, 2024]

This report outlines three complementary cybersecurity approaches (functional, lifecycle, and threat-based) that frontier AI developers and policymakers can use to assess how comprehensive their risk management practices are and address significant gaps. The authors recommend starting with a functional approach based on the NIST AI RMF.

Ee, S., O'Brien, J., Williams, Z., El-Dakhakhni, A., Aird, M., & Lintz, A. (2024). *Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach.* Institute for AI Policy and Strategy. http://arxiv.org/abs/2408.07933

Maturity models

The frameworks in this category provide systematic ways to assess how well organizations manage AI risks, using defined maturity levels (e.g., beginner, intermediate, advanced) and assessment criteria. They typically include scoring systems or rubrics and aim to help organizations understand their current strengths and weaknesses, and identify paths for improvement.

Framework to Rate AI Developers’ Risk Management Maturity [SaferAI, 2024]

This paper presents a methodology for rating or scoring the risk management maturity of frontier AI developers. The framework combines established risk management principles with AI-specific approaches (e.g., red teaming, risk thresholds), transforming them into a rating system that uses clear, quantitative criteria to evaluate how effectively AI developers implement risk management. The framework is narrowly focused on evaluating the model itself, rather than evaluating the model-in-deployment (the environment, stakeholders, and socio-technical ecosystem where the AI is actually used).

Campos, S., Papadatos, H., Roger, F., Touzet, C., & Murray, M. (2024). *A Framework to Rate AI Developers’ Risk Management Maturity*. SaferAI. https://www.safer-ai.org/research-posts/a-framework-to-rate-ai-developers-risk-management-maturity

Evolving AI Risk Management: A Maturity Model Based on the NIST AI Risk Management Framework [Dotan et al, 2024]

This preprint presents a foundation for a maturity model to evaluate how effectively organizations that manage and develop AI systems adhere to best practices in sociotechnical harm mitigation. The authors include a flexible questionnaire and scoring guidelines based on the standards put forward by NIST.

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., & Scarpino, J. (2024). Evolving AI risk management: A maturity model based on the NIST AI risk management framework. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2401.15229

Novel approaches

The frameworks in this category describe new methods specifically designed for frontier AI systems, addressing risks and features - such as rapid capability advancement and emergent behaviors - that existing risk management approaches cannot adequately handle.

Affirmative safety: An approach to risk management for high-risk AI [Wasil et al., 2024]

This paper proposes an approach to risk management called 'affirmative safety', in which those creating or deploying high-risk AI systems are required to demonstrate proof of safety prior to release. The authors outline four categories of evidence: technical, cognitive, developmental and operational. They also describe complementary practices – like robust information security or an established safety culture – that can support or strengthen an affirmative safety case.

Wasil, A. R., Clymer, J., Krueger, D., Dardaman, E., Campos, S., & Murphy, E. R. (2024). *Affirmative safety: An approach to risk management for high-risk AI*. In *arXiv [cs.CY]*. arXiv. http://arxiv.org/abs/2406.15371

Probabilistic Risk Assessment for AI [Centre for AI Risk Management & Alignment, 2024]

This paper introduces a systematic and generalized framework for adapting Probabilistic Risk Assessment (PRA) methods, used in high-reliability industries, to evaluate AI systems' risks. The framework introduces several methodological innovations, including risk pathway modeling, prospective risk quantification, and systematic analysis of both system capabilities and failures. It's implemented as a practical assessment workbook that integrates various assessment approaches to enable standardized risk evaluation across different AI systems.

Center for AI Risk Management & Alignment. (2024). *Probabilistic Risk Assessment for AI*. https://pra-for-ai.github.io/pra/

AI Hazards Management: A framework for the systematic management of root causes for AI risks [Schnitzer et al., 2023]

This paper introduces the AI Hazard Management (AIHM) Framework, a structured process to systematically identify and address the root causes of AI risk. The framework builds upon a preliminary list and original taxonomy of AI Hazards. The AI Hazards taxonomy describes when, how, and by whom to treat an AI hazard during the development and operation of an AI system, and is used by the authors to classify the preliminary list of AI Hazards.

Schnitzer, R., Hapfelmeier, A., Gaube, S., & Zillner, S. (2023). *AI Hazard Management: A framework for the systematic management of root causes for AI risks*. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2310.16727

Emerging practice

Frameworks in this category focus on documenting and sharing real-world practices in AI risk management, creating a knowledge base of current approaches and lessons learned. They typically compile examples from multiple organizations, providing insights into how different actors are addressing AI risks.

Emerging Processes for Frontier AI Safety [UK DSIT, 2023]

This UK Government document offers an idea bank or 'menu' of actual safety practices in use or under active consideration by developers of frontier AI systems, academia and broader civil society. Unlike the NIST AI RMF or ISO/IEC guidelines, this publication does not prescribe strict controls and instead compiles examples of "what good policy could look like". The document groups these examples by theme (e.g., government mechanisms, testing & evaluation, external engagement) but not by importance.

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