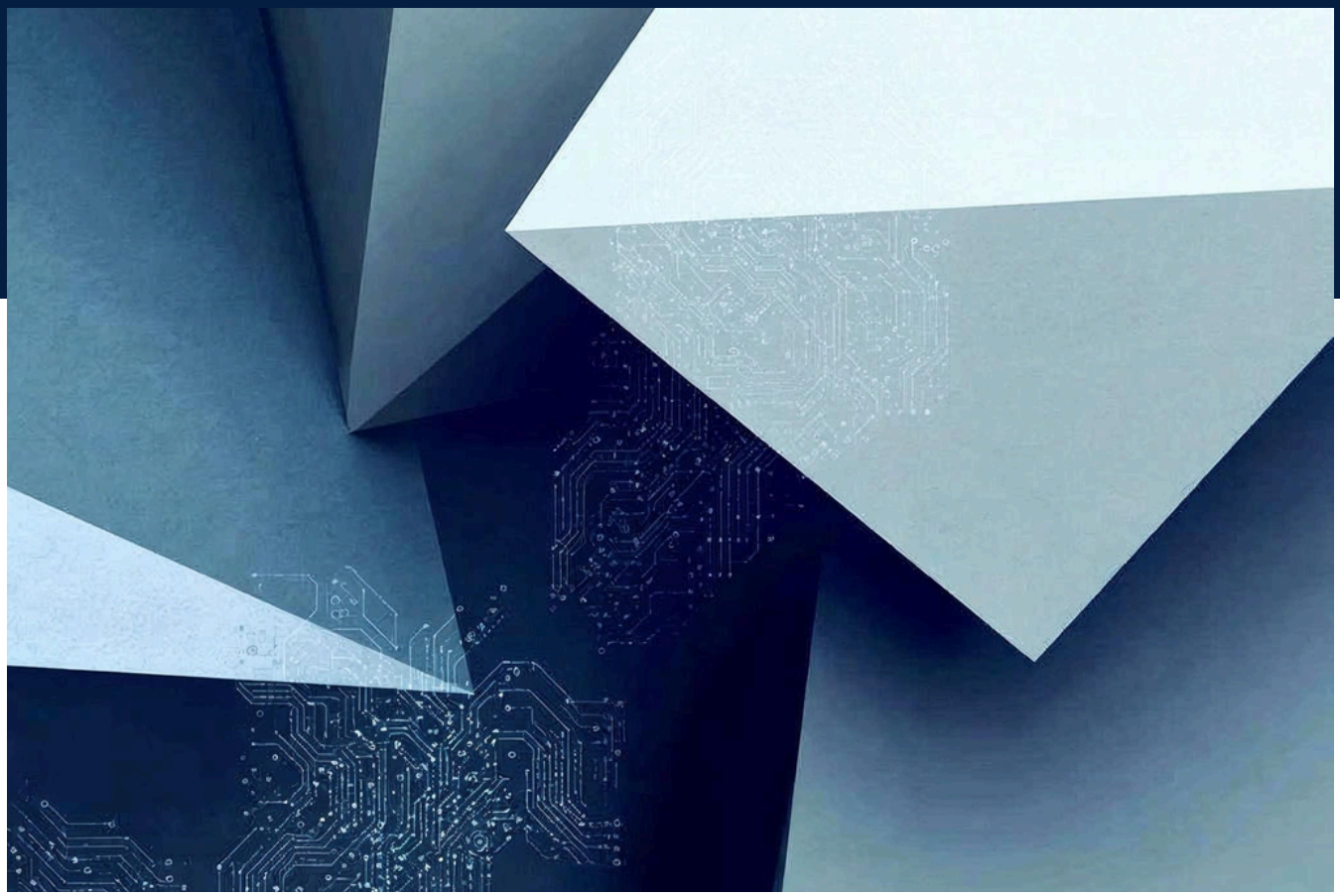
RESEARCH MEMO
2025 OCTOBER

# Safety Frameworks and Standards: A comparative analysis to advance risk management of frontier AI



Authors: Marta Ziosi, James Gealy, Miro Plueckebaum, Daniel Kossack, Simeon Campos, Lama Saouma, Uzma Chaudhry, Lisa Soder, Merlin Stein, Nicholas Caputo, Connor Dunlop, Jakob Mökander, Enrico Panai, Tom Lebrun, Charles Martinet, Ben Bucknall, Rebecca Weiss, Koen Holtman, Patricia Paskov, Saad Siddiqui, Fazl Barez, Ranj Zuhdi, Peter Slattery and Florian Ostmann

# Safety Frameworks and Standards: A comparative analysis to advance risk management of frontier AI

**Authors:** Marta Ziosi[1,], James Gealy[2], Miro Plueckebaum[1], Daniel Kossack[2], Simeon Campos[2], Lama Saouma[1,], Uzma Chaudhry[3], Lisa Soder[4], Merlin Stein[5], Nicholas Caputo[1], Connor Dunlop[6], Jakob Mökander[10, 11], Enrico Panai[7], Tom Lebrun[9], Charles Martinet[12, 1], Ben Bucknall[1,5], Rebecca Weiss[13], Koen Holtman[15], Patricia Paskov[1], Saad Siddiqui[8], Fazl Barez[1], Ranj Zuhdi[14], Peter Slattery[16], Florian Ostmann[3]

**Affiliations:** [1]Oxford Martin AI Governance Initiative, University of Oxford; [2]SaferAI; [3]AI Standards Hub; [4]Interface; [5]University of Oxford; [6]Ada Lovelace Institute; [7]Catholic University of the Sacred Heart (Milan); [8]Safe AI Forum; [9]Standards Council of Canada; [10]Tony Blair Institute for Global Change; [11]Yale Digital Ethics Center, Yale University; [12]Centre pour la Sécurité de l'IA (CeSIA); [13]MLCommons Association; [14]SDLC Compliance LTD; [15]AI Standards Lab; [16]MIT FutureTech, MIT.

**Abstract:**

This document systematically compares Frontier Safety Frameworks (FSFs) with international risk management standards, identifying areas of convergence, divergence, and opportunities for mutual reinforcement. FSFs have emerged rapidly in response to pressing governance needs, introducing concrete mechanisms such as capability thresholds, predefined frontier risks, evaluation triggers, and, in some cases, incident reporting channels and external assessments. While FSFs are agile and provide pragmatic, frontier-specific features, they often leave implicit fundamental considerations, including the definition of risk, the rationale for threshold selection, and the mapping of evaluation outcomes to risk severity and likelihood, to cite a few. By contrast, international risk management standards offer a mature vocabulary and structure, refined through consensus and validated through decades of application in high-stakes sectors. They emphasize assessing the adequacy of the overall risk management system, the justification of risk criteria, systematic and comprehensive processes, including for risk identification and analysis, clarity in linking assessment results to risk, and evaluation of the effectiveness of specific techniques, among others. However, these standards frequently remain abstract, were not designed with frontier AI in mind, and are not updated at a pace commensurate with its rapid development. The analysis concludes that integrating the systematic rigor of international standards with the frontier-specific innovations of FSFs offers a promising path toward more coherent, effective, and internationally harmonized practices. The document provides recommendations for the development of future standards for frontier AI risk management.

# 1. Introduction

The rapid advancement of frontier AI has highlighted the need for robust yet agile risk management practices, with several frontier labs developing Frontier Safety Frameworks (FSFs) outlining policies to address the critical risks from frontier AI (FMF, 2025). While it is common for a novel field to witness the development of distinct in-house practices and terminologies, it is positive for it to move toward more solid structures and harmonization (Yates & Murphy, 2019), which are currently lacking in FSFs (Pistillo, 2025). Risk management standards, developed over many years and tried and tested in fields such as medical devices, automated vehicles or aviation, provide well-developed terms and practices. However, AI risk management standards, differently from FSFs, were designed for AI systems prior to the "general-purpose AI" era and are not updated as regularly, making them less applicable to the dynamic and scalable nature of frontier AI (NIST, 2024).

Over the last two years, an increasing number of initiatives have developed proposals for frontier AI risk management. These have been facilitated both by government bodies (e.g., NIST RMF Generative Profile, EU GPAI Code of Practice) as well as by individual researchers (Barrett et al., 2025; Campos et al., 2025). These provide valuable blueprints advancing what frontier AI risk management could or should look like. Such proposals indirectly build both on risk management standards and frontier safety frameworks, but the relation between the two is often left implicit or used to advance proposals rather than for the sake of systematic analysis. To enhance the effectiveness and coherence of these proposals, enable the development of further risk management efforts on frontier AI and support greater international harmonisation, it is essential to make this relationship explicit and to articulate the added value it can provide. This memo asks, "*How do risk management standards and Frontier Safety Frameworks compare? And what are the implications for advancing frontier AI risk management?*".

Motivated by these considerations, this memo **aims to compare key components of Frontier Safety Frameworks to international Risk Management Standards**, identifying gaps, conflicts and similarities, and highlighting opportunities where one can learn from the other. By doing so, this document provides a set of key recommendations for the development of future risk management standards for frontier AI.

# 2. Background and Scope

Frontier AI refers to highly capable AI that could possess dangerous capabilities sufficient to pose severe risks to public safety (Anderljung et al., 2023). To manage such risks, companies developing frontier AI have introduced a set of Frontier Safety Frameworks (FSFs). These frameworks outline processes to define when AI systems reach high-risk capability thresholds and trigger specific risk assessments and mitigation measures (FMF, 2025). Both government and industry recognized the importance of FSFs through the Frontier AI Safety Commitments (2024) announced at the AI Seoul Summit in May 2024. Notwithstanding their

importance, FSFs remain a nascent concept; while they increasingly share a high-level structure, they often differ in terms of terminology, specific use of processes and implementation (FMF, 2024). This analysis considers the FSFs of Anthropic (2025), OpenAI (2025), Microsoft (2025), Google DeepMind (2025), and Meta (2025) as illustrative examples. Their selection is not intended as an assessment of quality, but rather reflects the fact that these frameworks are among the most established and longstanding frameworks to date.

International standards[1] are documents developed through consensus-building processes, approved by recognized bodies, and reviewed and updated every few years[2] (ISO/IEC Guide 2:2004). They encode best practices that have been developed over decades, and provide common and repeatable rules and guidelines that govern activities and their outcomes. The focus of this analysis is specifically on risk management standards, such as ISO 31000, and its application to AI, ISO/IEC 23894. It also considers standards related to safety, including ISO/IEC Guide 51 on the inclusion of safety aspects in standards, as well as domain-specific standards, such as ISO 14971:2019 on the application of risk management to medical devices, insofar as these enrich the comparison with Frontier Safety Frameworks (FSFs) and their relevance for frontier AI. Closely related standards on management systems (e.g., ISO/IEC 42001 for AI) are not the focus of this analysis. These govern how an organisation runs its processes overall in order to achieve its quality objectives, of which risk management constitutes only a component. Although they bear relevance to some of the themes discussed here, they are not examined in detail so as to maintain a clear and targeted focus for the analysis. We refer to ISO 31073:2022 Risk management — Vocabulary to provide the relevant definitions across the document.

# 3. Frontier Safety Frameworks & Risk Management Standards: A Comparison

FSFs and risk management standards have a lot in common when it comes to agreeing on and producing a set of rules and guidelines to manage risks from frontier AI. However, they differ on a conceptual and operational level.

On the one hand, FSFs share common elements (Buhl et al., 2025; FMF, 2024; METR, 2025), such as governance and transparency measures, a set of critical risks that are to be assessed against specific thresholds, methods to evaluate such risks, and decisions around deployment and mitigation strategies. Risk management standards, on the other hand, generally consist of a risk management system (or framework) - which requires outlining the organizational policy regarding risk management - and a risk management process - which aims to achieve compliance with the policy and includes risk identification, risk analysis, risk evaluation and risk treatment.

[1] This document will focus primarily on ISO and IEC standards for simplicity.
[2] ISO standards are reviewed and updated every 5 years, while IEC standards review period could range from 3 to 12 years.

In the following sections, we compare elements of risk management standards to those of FSFs and propose a set of recommendations to guide the development of future frontier AI risk management standards. Below is an overview table summarising the key takeaways.

**Key takeaways for frontier AI risk management standards:**

**Risk Management System:**

- **Clarify roles and responsibilities:** Ensure top management holds ultimate responsibility, with clear separation of managerial and oversight functions, aligned with a company's specific governance structures, and supported with adequate resources.
- **Foster reporting practices and transparency:** Create effective internal and external reporting mechanisms, including protected channels for reporting non-compliance and for incidents, with formal escalation processes for serious cases. Commit to external disclosures (e.g., risk assessment results), with clear rules for redactions.
- **Integrate external expertise and independent assessments:** Define clear triggers for involving external experts in the risk management process. Clearly define the commitment to and extent of external involvement, including for external testing and assessments.
- **Embed regular updates and continuous improvement:** Set clear criteria for when the risk management system should be updated and include mechanisms for regular review. Go beyond updates by committing to continuous improvement of the system's suitability, adequacy, and effectiveness.

**Risk Criteria:**

- **Define scope and context:** Incorporate explicit considerations of scope and context to tailor the risk assessment process to the expected application(s) of the model and its use.
- **Justify risk criteria:** Provide clear rationales for the choice of risk criteria or thresholds. Reference existing approaches to criteria or standard thresholds from specific fields where available. Include how uncertainties are addressed and why qualitative or quantitative approaches are used.
- **Link thresholds to risk:** Ensure thresholds are explicitly connected to risk as a function of severity and likelihood. Establish consistency across tiers or levels so that thresholds can be meaningfully compared and interpreted as corresponding risk levels.

**Risk Identification:**

- **Defining risk:** specify the definition of risk on which the process of risk identification is based (e.g., if marginal, describe how it is defined and in reference to which baseline) and the elements considered in risk identification (e.g., sources, potential events and outcomes, possible impacts, etc.). Establish a set of criteria for

determining which risks fall within scope, such as plausibility, severity, measurability, and immediacy. .

- **Comprehensive and structured risk identification:** explicitly include how the process of risk identification reliably identifies known risks, how it accounts for the identification of new, emergent risks and how uncertainty is accounted for in the risk identification process. Include the involvement of any external experts or stakeholders and how this enriches the risk identification process.
- **Specify and justify use of techniques:** Explicitly link the techniques employed (e.g., red-teaming, threat modelling) to their specific role and value for risk identification (e.g., distinguishing them from their use in mitigation or evaluation). Provide a justification for how identified risks are derived from such methods.

**Risk Analysis:**

- **Depth and breadth of analysis:** Define clear triggers or criteria for when deeper or more comprehensive analysis is required, ensuring prioritization across different risk domains.
- **Diversity of techniques:** Employ a combination of relevant methods beyond just model evaluations for a more comprehensive assessment and to better address uncertainty and complexity. Where comparable models are used, systematically assess how differences may affect risk occurrence.
- **Effectiveness of controls:** Explicitly assess the adequacy of existing mitigations as part of the analysis.
- **Risk estimation:** Establish a transparent link between analysis results and risk criteria, translating outcomes into severity and likelihood assessments, and clarify how results from multiple techniques are aggregated.

**Risk Evaluation:**

- **Design an iterative process:** Risk evaluation should be treated as an iterative process that goes beyond a single decision point, informing both the selection of mitigations and subsequent deployment choices at relevant evaluation points (e.g., pre- and post-mitigations).
- **Determine clear options upfront:** Specify the concrete options that risk evaluation can trigger (e.g., proceed, strengthen mitigations, stop development, withhold release), ensuring decisions are systematically tied to evaluation outcomes.
- **Establish explicit protocols:** Clarify who participates in the evaluation process, including the role of internal management and external experts, and how their input informs final decisions.
- **Record and communicate:** Ensure that evaluations are documented, communicated, and validated at the appropriate organizational level. Make explicit what information is considered, how it is compared to criteria, and how uncertainties or insufficient evidence are handled.

**Risk Treatment:**

- **Define and justify mitigation plans:** Specify which mitigations apply to which levels of risk, justify their selection, and outline how they will be implemented.

- **Ensure robustness and transparency in application:** Report how mitigations are applied in practice, how their effectiveness and sufficiency are assessed, and which methods are used to establish this.
- **Address residual risk explicitly:** Include procedures for assessing residual risk after mitigations, and clarify how remaining risks are weighed against expected benefits in deployment decisions, if at all.

## 3.1 Risk management system

In risk management standards, the risk management system[3] entails defining responsibilities, designing and implementing the plan for risk management, and ensuring its evaluation and review. FSFs incorporate various elements in establishing a risk management framework, often linking them to broader aspects of risk governance. These can include aspects such as allocation of responsibilities, internal communication and decision-making, transparency and external input.

Regarding setting responsibilities, most risk management standards, including ISO 31000 and ISO/IEC 23894, place ultimate responsibility and accountability with top management. This ensures that those setting the priorities of the organization have cognizance of the risks. They also draw a difference in function between top management and oversight bodies; where top management is accountable for managing risk while oversight bodies are accountable for overseeing risk management. In FSFs, responsibility for managing risks is often assigned to a specific actor (e.g., Anthropic Responsible Scaling Officer, Microsoft's Executive Officers) or groups of actors (e.g., OpenAI's Safety Advisory Group, Google AGI Safety Council). Decisions usually have to go through some higher levels of management. Anthropic, for example, requires the Responsible Scaling Officer to make the ultimate determination on the sufficiency of some measures with the CEO, and share plans with the Board of Directors and the Long-Term Benefit Trust. Some FSFs also hint towards an oversight function. OpenAI's Preparedness Framework, for example, cites the role of the Safety and Security Committee (SSC) of the OpenAI Board of Directors as having an oversight role.

Different FSFs differ in how specifically they define these relations and to the kinds of actors they assign these different roles to (e.g., whether they are one person or a group, whether they are the CEO, the leadership team or another kind of body). These variations may also be driven by the differences between these companies, where some make Frontier AI their main product (e.g., Anthropic, OpenAI), while others develop it alongside an array of other products and within a wider corporate structure (e.g., Google, Microsoft, Meta). There is currently little guidance on how to adapt responsibility allocation for frontier AI to different corporate structures, both for the cases cited above and for cases where the company

[3]Some standards, such as ISO/IEC 31000, refer to risk management activities, recommendations and requirements encompassing the risk management process as the "risk management framework". Other standards, like some based on ISO/IEC Guide 51, refer to this as the risk management system. This document uses risk management system.

developing the frontier AI model is also the one deploying it into AI systems. At a minimum, risk management standards stress the importance of allocating sufficient resources across the firm to ensure that risk management can be carried out. This aspect is left ambiguous or not explicitly addressed in most FSFs, but is key for ensuring that organizational efforts can be effectively implemented by the designated responsible actors.

Risk management standards like ISO 31000 and ISO/IEC 23894 describe processes for establishing communication and consultation. Communication refers to the dissemination of information to targeted audiences, whereas consultation entails engaging participants in providing feedback with the expectation that it will inform and influence decisions or related activities. FSFs usually include aspects on external accountability as well as on the involvement of external experts. Most FSFs incorporate a section on transparency, which may include, for example, the public release of model cards (e.g., Meta), the rationale underlying deployment decisions, and the disclosure of certain information on capability evaluations and implemented safeguards when a model exceeds a high-risk threshold (e.g., OpenAI). The extent and granularity of disclosure remains discretionary, with some FSFs noting that information may be redacted or summarized where necessary, for instance to safeguard intellectual property or security (e.g., OpenAI).

FSFs often mention establishing channels for reporting instances of non-compliance, such as Microsoft reporting channels, OpenAI raising concerns policy and Anthropic's non-compliance notifications. Importantly, some FSFs use these channels for serious incident reporting (e.g., Microsoft) or refer separately to the establishment of incident reporting procedures (e.g., Anthropic ASL2 Security Standard). This is a key aspect in managing risks from Frontier AI and one that is not explicitly addressed in general risk management standards such as ISO 31000 and ISO/IEC 23894, nor in ISO/IEC Guide 51 on safety aspects.

FSFs also tend to mention the consultation of external actors. However, they differ in the triggers for it and in the extent of external involvement. In FSFs, external consultation can be referred to as input and feedback from external actors such as experts (e.g., Anthropic, Microsoft, to cite a few), as well as "third-party evaluations" and "third-party stress-testing" (e.g., OpenAI). Given the risks from Frontier AI, the latter is a crucial aspect to manage such risks (Bucknall & Trager, 2023; Shevlane, 2022) and is rarely featured in risk management standards. FSFs also often include a section on future updates and revisions, which is a required step in risk management standards. Importantly, risk management standards stress that an organization should continually improve the suitability, adequacy and effectiveness of the risk management framework. To that effect, FSFs should not only include updates but also refer to how these latter aspects are ensured.

**Key takeaways for frontier AI risk management standards:**

- **Clarify roles and responsibilities:** Ensure top management holds ultimate responsibility, with clear separation of managerial and oversight functions, aligned with a company's specific governance structures, and supported with adequate resources.
- **Foster reporting practices and transparency:** Create effective internal and external reporting mechanisms, including protected channels for reporting non-compliance and for incidents, with formal escalation processes for serious cases. Commit to external disclosures (e.g., risk assessment results), with clear rules for redactions.
- **Integrate external expertise and independent assessments:** Define clear triggers for involving external experts in the risk management process. Clearly define the commitment to and extent of external involvement, including for external testing and assessments.
- **Embed regular updates and continuous improvement:** Set clear criteria for when the risk management system should be updated and include mechanisms for regular review. Go beyond updates by committing to continuous improvement of the system's suitability, adequacy, and effectiveness.

.

## 3.2 Risk criteria

In risk management standards, risk criteria refer to the terms of reference by which an organization evaluates the significance of the risks that they identify and make decisions concerning risks (ISO 31073:2022). In most FSFs, risk criteria are operationalized as risk or capability thresholds, which are levels of an AI system's performance or capabilities that, when reached, require implementation of specific mitigation measures (Campos et al., 2025). FSFs establish different thresholds for the different risks in scope, such as for CBRN, Cyberoffense, and AI R&D, to cite a few.

Risk management standards often treat risk criteria as part of the initial stage in defining the scope and context of risk management.[4] This stage typically precedes the risk assessment process and involves consideration of the intended purpose of the AI system as well as the internal and external context of its application. The objective is to adapt the risk management process to the specific environment in which it will be implemented. As FSFs are primarily concerned with general-purpose models that, by design, are not tied to a single application context, the manner in which they should establish considerations of scope and context remains uncertain. These considerations may be applicable to a company's internal deployments where the scope and context of application are well defined and which themselves may give rise to severe risks (Stix et al., 2025). For models that are publicly deployed into different systems, information on external uses of such models can inform scope and context considerations through contractual arrangements or techniques such as post-deployment monitoring (Stein et al., 2024). In this regard, standards such as ISO/IEC

[4]Depending on the standard, criteria can be defined separately from scope. ISO 31000 puts them together, but for product safety it is about the intended use of the system and the risk acceptability criteria are separate yet connected.

23894 prescribe that organizations collect and review publicly available information on comparable systems in the market.

Regarding risk criteria more specifically, risk management standards (e.g., ISO/IEC 23894) prescribe that, when establishing these, organizations should consider factors such as the how consequences (both positive and negative) and likelihood will be defined and measured, how the level of risk is to be determined, how consistency in measurement will be ensured, and how combinations and sequences of multiple risks will be taken into account, among others. In FSFs, thresholds tend to be informed by or shaped around the specific capabilities (or occasionally outcomes, in the case of Meta) that are related to the specific risks in scope (e.g., cyber capabilities for cyberoffense risks). By relying on thresholds for specific risks, FSFs tend to rely on a direct application of possible risk acceptance criteria, and it remains unclear which other factors are considered in shaping these thresholds. There also tends to be a lack of explicit justification in the choice of specific risk thresholds and in their use and measurement. Anthropic is one of the most explicit by explaining that its thresholds, called "AI Safety Levels" (ASL), are modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials (Anthropic, 2025). The extent to which uncertainties inform the determination of thresholds is rarely made explicit, although such considerations may occur internally.

Risk management standards in several high-risk industries tend to rely on quantitative thresholds (Campos et al., 2025). For instance, the aviation industry adheres to an acceptable frequency of catastrophic accidents, defined as "failure conditions that would prevent continued safe flight and landing", limiting their occurrence to less than one per billion flight hours (Campos et al., 2025). This corresponds to approximately one catastrophic event per 114,155 aircraft operating years (Federal Aviation Administration, 1988). Thresholds in FSFs tend to be qualitative (Caputo et al., 2025). Capability thresholds are often expressed through qualitative scenarios that describe a catastrophic situation to be avoided (e.g., see Anthropic, Microsoft, OpenAI, to cite a few). While risk criteria tend to rely on a higher-level principle (e.g., number of deaths over time), FSFs feature different thresholds for different types risks (e.g., CBRN-4 is different from AI R&D-4), and different operationalisation for the same risks across companies. Some rely on just one threshold per risk domain (e.g., Amazon), while others have multiple "tiers" or levels (e.g., "high" and "critical" in OpenAI).

Whereas risk criteria in risk management standards are articulated through considerations directly linked to risk (e.g., consequences and likelihood), thresholds in FSFs tend to be only implicitly connected to risk. Capabilities and their associated mitigations are used to delineate acceptable levels of risk. However, there is often a lack of explicit reference or mapping to likelihoods, consequences, or the methods by which these are determined. Moreover, it remains unclear whether the different tiers established within FSFs are internally consistent or comparable with one another (Pistillo, 2025); for example, whether they correspond to equivalent risk levels, understood as combinations of severity and probability or not.

**Key takeaways for frontier AI risk management standards:**

- **Define scope and context:** Incorporate explicit considerations of scope and context to tailor the risk assessment process to the expected application(s) of the model and its use.
- **Justify risk criteria:** Provide clear rationales for the choice of risk criteria or thresholds. Reference existing approaches to criteria or standard thresholds from specific fields where available. Include how uncertainties are addressed and why qualitative or quantitative approaches are used.
- **Link thresholds to risk:** Ensure thresholds are explicitly connected to risk as a function of severity and likelihood. Establish consistency across tiers or levels so that thresholds can be meaningfully compared and interpreted as corresponding risk levels.

## 3.3 Risk identification

In risk management standards, risk identification entails finding, recognizing and describing risks that are reasonably foreseeable for assessment (ISO 31073:2022). In such standards, risk identification does not specify a set of risks, but rather a set of elements or procedures that help identify risks. FSFs do not typically detail out risk identification as a process; rather, they tend to specify a predefined set of risks requiring assessment. These are usually framed as catastrophic or "frontier" risks, such as CBRN risks or cyberoffense.

Risk is defined differently in different risk management standards. For some, "risk" refers to the "effect of uncertainty on objectives" (ISO 31000), for others it refers to the "combination of the probability of occurrence of *harm* and the *severity* of that *harm*" (ISO 14971:2019) FSFs tend to leave implicit their taken definition of risk. OpenAI explicitly refers to "marginal risk", the risk a model poses compared to a baseline. Other FSFs also seem to indirectly imply a reliance on marginal risk (e.g., Microsoft applying mitigations "so that the marginal risks a model may pose are appropriately addressed", p.7, 2025). Concurrently, some FSFs are explicit with regard to the criteria they consider for a risk to be in scope. In the case of OpenAI, for example, a risk needs to be plausible, measurable, severe, net new and instantaneous or irremediable in order to be considered. Similarly, for Meta risk needs to be plausible, catastrophic, net new and instantaneous or irremediable.

Risk management standards have an emphasis on risk identification being comprehensive, covering various aspects such as sources of risk (e.g., data, physical settings, operations, and practices), potential events and outcomes, possible impacts (on individuals, communities, groups, and vulnerable populations), limitations of knowledge and reliability of information; to cite a few (ISO 31000). Several FSFs mention a "holistic process" when referring to their process of risk assessment in general (e.g., OpenAI, Microsoft). However, the meaning of "holistic" is not clearly operationalised or articulated, and many frameworks focus primarily on analysing a predefined set of risks (e.g., Anthropic, Google). Some FSFs explicitly

emphasize the importance of identifying emerging risks, for example through the use of threat modeling (e.g., Meta), which employs threat scenarios to illustrate how different actors might exploit a frontier AI model to produce catastrophic outcomes. Meta also mentions consultation and workshops with external experts for that purpose, hinting at a more comprehensive risk identification process. Additionally, even though the kinds of cutting-edge risks covered in FSFs might be subject to a high-level of uncertainty, FSFs tend not to make any mention of limitations of knowledge and reliability of information when characterising such risks, a feature which is present in risk management standards.

Regarding specific methods for risk identification, risk management standards tend to be high-level and not delve into specific techniques. Yet, specific risk identification techniques are cited in standards like ISO 31010 (risk assessment techniques), IEC 61025 (fault tree analysis), and IEC 61882 (hazard and operability studies, or HAZOP). These provide more detailed guidance on how to select and apply the appropriate techniques based on the situation. Their applicability to frontier AI risks, however, remains only little explored (Koessler & Schuett, 2023), especially for discovering emerging frontier AI risks. FSFs mention several techniques that are relevant for frontier AI risk identification, such as red-teaming and threat modelling. The specific way in which techniques are applied and used for risk identification, however, is often left general or undisclosed. For instance, red-teaming exercises are frequently cited as being used for uncovering new risks, yet they are also commonly referenced as a method for testing the robustness of mitigations (e.g., Meta). While it is normal for a technique to serve multiple purposes, it remains important to make explicit how it is specifically applied in the context of risk identification (e.g., scenario-driven threat modelling to identify possible negative outcomes enabled by a model).

**Key takeaways for frontier AI risk management standards:**

- **Defining risk:** specify the definition of risk on which the process of risk identification is based (e.g., if marginal, describe how it is defined and in reference to which baseline) and the elements considered in risk identification (e.g., sources, potential events and outcomes, possible impacts, etc.). Establish a set of criteria for determining which risks fall within scope, such as plausibility, severity, measurability, and immediacy.
- **Comprehensive and structured risk identification:** explicitly include how the process of risk identification reliably identifies known risks, how it accounts for the identification of new, emergent risks and how uncertainty is accounted for in the risk identification process. Include the involvement of any external experts or stakeholders and how this enriches the risk identification process.
- **Specify and justify use of techniques:** Explicitly link the techniques employed (e.g., red-teaming, threat modelling, fault tree analysis) to their specific role and value for risk identification (e.g., distinguishing them from their use in mitigation or evaluation). Provide a justification for how identified risks are derived from such methods.

## 3.4 Risk analysis

In risk management standards, risk analysis is the process to comprehend the nature of risk and to determine the level of risk (ISO 31073:2022). It provides the basis for risk evaluation[5] and decisions about risk treatment. It is an in-depth process that involves thoroughly examining uncertainties, sources of risk, potential consequences, likelihood, possible events and scenarios, existing controls, and how effective those controls are. In FSFs, such analysis is often carried out through model evaluations, sometimes accompanied by some amount of threat modelling and forecasting.

Whereas high-level risk management standards generally do not prescribe specific risk analysis techniques (e.g., ISO 31000),[6] FSFs primarily emphasize model evaluations as a specific technique. While heavily focused on evaluations, different FSFs may describe their use at different stages and at different levels of depth. For instance, Anthropic performs an initial assessment of relevant capabilities, followed, where appropriate, by a more comprehensive evaluation. Similarly, Microsoft assesses for leading indicators of high-risk capabilities that, if identified, trigger a more in-depth capability assessment. The use of triggers for initiating more extensive evaluations enables more efficient assessment and prioritization across diverse frontier AI risks. Such mechanisms are rarely reflected in general risk management standards. For instance, ISO/IEC Guide 51, section 6.1, provides explicit and detailed guidance on the process of risk analysis but does not address practices such as the preliminary identification of leading indicators, which are incorporated into certain FSFs.

Risk management standards acknowledge the complexities involved in analyzing highly uncertain events and recommend the use of multiple complementary techniques to generate more comprehensive insights (e.g., ISO 31000, ISO/IEC 23894). FSFs rely on techniques for analysis beyond just evaluations. However, there tends to be less of a focus on these. For example, although several FSFs explicitly reference threat modeling, the extent and depth of its use remain unclear, and the resulting threat models are rarely disclosed. Risk management standards such as ISO 31000 highlight the importance of assessing the effectiveness of existing controls as part of the analysis. Some FSFs consider it (e.g., Meta, Anthropic, OpenAI), but this should be a consistent requirement across.

Risk management standards emphasize the use of the best available information, supplemented by additional inquiry where necessary, and underscore the need to ensure that the results of the analysis are both reliable and representative of the relevant context (ISO 31000). For some specific standards, this includes considering relevant information from an assessment of a similar device, like in the case of ISO 14971:2019 on the application of risk management to medical devices. The practice of drawing on information from similar models is also evident in some FSFs. For example, Meta uses the concept of comparable models to

[5] The use of the term risk evaluation in risk management standards is different from model evaluations as it is used in FSFs. In risk management standards, risk evaluation describes the process of comparing the results of risk analysis with risk criteria to determine whether the risk is acceptable or tolerable (ISO 31073:2022).
[6] That said, standards such as ISO 31010 provide detailed methodologies also for risk analysis though, as mentioned before, their applicability to frontier AI has not yet been systematically examined.

use as a reference class for which evaluations to run. Anthropic uses the concept of notably more capable models to decide whether a more comprehensive assessment is needed. In relevant risk management standards, the use of an existing risk analysis from a similar device needs to be based on a systematic evaluation of the effects that the differences can have on the occurrence of hazardous situations (ISO 14971:2019). Whether this is conducted for FSFs when relying on similar models is unclear or not explicitly communicated.

FSFs show other specific attempts to account for context in the analysis in novel ways, previously unseen in risk management standards on AI. In running their evaluations, for example, some seek to equip the model with appropriate scaffolding and other augmentations to make it more likely that they are also assessing the capabilities of systems that will likely be produced with the model (e.g., Google). Others design their evaluations to account for the deployment context of the model. This includes assessing whether risks will remain within defined thresholds once a model is deployed or released using the target release approach (e.g., Meta).

According to risk management standards, results from risk analysis should translate into likelihood or severity assessments to inform risk evaluation. In that respect, a core aspect expressed in risk management standards is risk estimation, where the level of risk is finally measured. As mentioned above, the strong reliance on evaluations in FSFs tends to leave unclear the relation between evaluation results and how they translate to an assessment of risk. Additionally, where FSFs rely on multiple analysis techniques, it is often left implicit how multiple results are aggregated and how or the extent to which they count towards the evaluation of the level of risk. Overall, there is a need for an explicit link between testing results and risk criteria as well as translatability to likelihood and severity assessments of real-world risks.

**Key takeaways for frontier AI risk management standards:**

- **Depth and breadth of analysis:** Define clear triggers or criteria for when deeper or more comprehensive analysis is required, ensuring prioritization across different risk domains.
- **Diversity of techniques:** Employ a combination of relevant methods beyond just model evaluations for a more comprehensive assessment and to better address uncertainty and complexity. Where comparable models are used, systematically assess how differences may affect risk occurrence.
- **Effectiveness of controls:** Explicitly assess the adequacy of existing mitigations as part of the analysis.
- **Risk estimation:** Establish a transparent link between analysis results and risk criteria, translating outcomes into severity and likelihood assessments, and clarify how results from multiple techniques are aggregated.

## 3.5 Risk evaluation

In risk management standards, risk evaluation involves comparing the results of the risk analysis with the established risk criteria to determine where additional action is required (ISO 31073:2022). In FSFs, this is often expressed by measuring where the capability levels lie with respect to the thresholds set. In FSFs, the name "risk evaluation" can be easily conflated with the term "model evaluations" described in the previous paragraph, which belongs to the step of risk analysis.

In risk management standards, risk evaluation informs a broad range of possible actions. Outcomes may include deciding to take no further action, exploring risk treatment options, conducting additional analysis to obtain greater clarity, or maintaining existing controls, among other possibilities (e.g., ISO 31000). FSFs tend to be rather pragmatic in their use of risk evaluation, often employing it to determine which mitigations to adopt and to guide deployment decisions. For instance, Meta directly links risk thresholds to the options of "stop development," "do not release," or "release." Similarly, Anthropic aligns each capability threshold with a corresponding set of deployment and security standards, which determine the safety and security measures to be implemented depending on the results of risk evaluation. Guide 51, addressing safety aspects in standards, emphasizes that this step should be iterative: risk should be evaluated both before and after the application of mitigations, thereby distinguishing the decision on which mitigations to deploy from subsequent deployment decisions (Section 6.1). This distinction is less common, but does appear in some FSFs. For example, Microsoft compares risk evaluations conducted before and after the application of mitigations to inform its deployment decisions. Although relatively rare, some FSFs are also explicit about the option of conducting further analysis. Anthropic, for instance, specifies that additional assessment may be undertaken if the organization cannot demonstrate that risk levels fall below a defined threshold.

In risk management standards, risk evaluation is not intended to be a purely formal or narrow calculation; rather, it should take into account the broader context as well as the actual and perceived consequences for both internal and external stakeholders. Some FSFs incorporate external actors into the evaluation process, either in decision-making or in reviewing outcomes. For example, Anthropic reports that it solicits both internal and external expert feedback on its capability report. Moreover, risk management standards stipulate that the outcomes of risk evaluation must be documented, communicated, and validated at appropriate organizational levels. While several FSFs require that risk evaluations be reviewed by senior levels of the organization, they rarely provide details on the evaluation process itself (e.g., what information is considered, how it is compared against thresholds), even though such information may be recorded internally.

**Key takeaways for frontier AI risk management standards:**

- **Design an iterative process:** Risk evaluation should be treated as an iterative process that goes beyond a single decision point, informing both the selection of mitigations and subsequent deployment choices at relevant evaluation points (e.g., pre- and post-mitigations).
- **Determine clear options upfront:** Specify the concrete options that risk evaluation can trigger (e.g., proceed, strengthen mitigations, stop development, withhold release), ensuring decisions are systematically tied to evaluation outcomes.
- **Establish explicit protocols:** Clarify who participates in the evaluation process, including the role of internal management and external experts, and how their input informs final decisions.
- **Record and communicate:** Ensure that evaluations are documented, communicated, and validated at the appropriate organizational level. Make explicit what information is considered, how it is compared to criteria, and how uncertainties or insufficient evidence are handled.

## 3.6 Risk treatment

In risk management standards, risk treatment refers to the process of modifying risk (ISO 31073:2022). During risk treatment, strategies for managing the risks identified during risk analysis are chosen. In FSFs, risk treatment is typically expressed in terms of "mitigations" or "safeguards". Mitigations are triggered either by detecting the presence of specific capabilities or indicators thereof, by risk levels or both.

In risk management standards, risk treatment may involve, among other measures, eliminating the source of risk, reducing its likelihood, or mitigating its potential consequences (e.g., ISO 31000). FSFs tend to describe mitigations in terms of whether they are safety or security mitigations or, alternatively, referring to them as deployment or security safeguards (e.g., Anthropic). Some FSFs prescribe upfront a set of mitigations for a specific level of risk. For example, Anthropic links each of its AI Safety Levels (ASLs) to a defined set of mitigations. A similar practice is adopted by Meta.

FSFs present a host of mitigations applicable to frontier AI that are rarely found in AI risk management standards like ISO/IEC 23894. Examples are defense in-depth, fine-tuning, misuse filtering and response protocol, and staged release to prepare the external ecosystem, to cite a few. Some standards related to safety, such as ISO/IEC Guide 51, advance some relatively detailed, yet still high-level proposals which can be used as general guidance (e.g., a three-step method for the reduction of risk which includes, in order of priority, inherently safe design; guards and protective devices; information for end users).

Risk management standards require a justification of risk treatment options and a risk treatment plan to specify how the chosen treatment options will be implemented. Risk management provides a structured approach to this process, which includes steps such as

developing and selecting suitable risk treatment options, planning and implementing the chosen treatments, evaluating their effectiveness, determining if the remaining (residual) risk is acceptable, and, if it is not, undertaking additional treatment measures (e.g., ISO 31000). While they may do so internally, FSFs report which host of mitigations they may apply, yet they rarely elaborate on how they specifically plan to implement them. OpenAI is a rare example that explicitly outlines a plan for safeguard selection as well as for establishing the sufficiency of mitigations. Once mitigations have been applied, some FSFs also refer to residual risk. For example, Meta includes an assessment of residual risk after mitigations have been applied to inform deployment decisions. Microsoft refers to the marginal benefits of a model outweighing any residual risk as one of the criteria to consider before deployment.

**Key takeaways for frontier AI risk management standards:**

- **Define and justify mitigation plans:** Specify which mitigations apply to which levels of risk, justify their selection, and outline how they will be implemented.
- **Ensure robustness and transparency in application:** Report how mitigations are applied in practice, how their effectiveness and sufficiency are assessed, and which methods are used to establish this.
- **Address residual risk explicitly:** Include procedures for assessing residual risk after mitigations, and clarify how remaining risks are weighed against expected benefits in deployment decisions, if at all.

# 4. Conclusion

The comparison between Frontier Safety Frameworks (FSFs) and international risk management standards highlights both convergence and divergence in how risks from frontier AI are conceptualized and addressed. FSFs have emerged rapidly to fill an urgent need, offering concrete mechanisms such as capability thresholds, predefined “frontier” risks, and analysis and mitigation measures specific to frontier AI. Risk management standards, by contrast, provide mature, structured approaches grounded in decades of practice across high-stakes industries, emphasizing comprehensiveness, justification of criteria, systematic processes, and continuous improvement.

The analysis shows that FSFs bring innovations tailored to the unique challenges of frontier AI, including incident reporting channels, triggers for evaluations, and external assessments, among others, which are not commonly found in traditional standards. They also tend to be pragmatic, operationalizing risk through capability thresholds tied to specific mitigation and deployment decisions. In doing so, FSFs provide concreteness and immediacy, but often leave implicit fundamental considerations, such as the definition of risk, the rationale for chosen thresholds, and how evaluation results map onto severity and likelihood of risks.

International risk management standards, by contrast, provide a mature vocabulary and structure, refined through consensus and tested through years of application in specific sectors. They emphasize comprehensiveness in processes such as risk identification and analysis, requiring attention to sources of risk, potential outcomes, uncertainties, and the reliability of information. Standards also stress systematic approaches, ensuring that responsibilities are clearly defined, risk evaluations are recorded and validated, and mitigations are justified. Unlike FSFs, however, they often remain at a high level of abstraction, are designed for application-specific systems before frontier AI, and are not updated with the frequency required to keep pace with frontier AI's rapid development.

Ultimately, the integration of the structured, systematic approach of risk management standards with the frontier-specific innovations of FSFs offers a promising path toward more coherent, effective, and internationally harmonized practices. Such an alignment would strengthen the governance of frontier AI, enhance trust and accountability, and better equip organizations and regulators to address the profound challenges posed by increasingly capable AI systems through the advancement of frontier AI risk management. This analysis represents an initial step in that direction and highlights the need for further research, particularly on how convergence between risk management standards and FSFs might be achieved in specific stages of the risk management process.

# Bibliography

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S.,

Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (No. arXiv:2307.03718). arXiv. https://doi.org/10.48550/arXiv.2307.03718

Anthropic. (2025). *Responsible Scaling Policy*. https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.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* (No. arXiv:2506.23949). arXiv. https://doi.org/10.48550/arXiv.2506.23949

Bucknall, B., & Trager, R. (2023). *Structured access for third-party research on frontier AI models:…*. Oxford Martin School. https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements

Buhl, M. D., Bucknall, B., & Masterson, T. (2025). *Emerging Practices in Frontier AI Safety Frameworks* (No. arXiv:2503.04746). arXiv. https://doi.org/10.48550/arXiv.2503.04746

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* (No. arXiv:2502.06656). arXiv. https://doi.org/10.48550/arXiv.2502.06656

Caputo, N., Campos, S., Casper, S., Gealy, J., Hung, B., Jacobs, J., Kossack, D., Lorente, T., Murray, M., Ó hÉigeartaigh, S., Oueslati, A., Papadatos, H., Schuett, J., Wisakanto, A. K., & Trager, R. (2025). Risk Tiers: Towards a Gold Standard for Advanced AI. *Oxford Martin AIGI*.

https://aigi.ox.ac.uk/publications/risk-tiers-towards-a-gold-standard-for-advanced-ai/

Federal Aviation Administration. (1988). *System Design and Analysis*. https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_25.1309-1A.pdf

FMF. (2024, November 8). Issue Brief: Components of Frontier AI Safety Frameworks. *Frontier Model Forum*. https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/

FMF. (2025, April 22). Introducing the FMF's Technical Report Series on Frontier AI Frameworks. *Frontier Model Forum*. https://www.frontiermodelforum.org/updates/introducing-the-fmfs-technical-report-series-on-frontier-ai-safety-frameworks/

*Frontier AI Safety Commitments, AI Seoul Summit 2024*. (2024). GOV.UK. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

Google DeepMind. (2025). *Frontier Safety Framework 2.0*. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0.pdf

Koessler, L., & Schuett, J. (2023). *Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries* (No. arXiv:2307.08823). arXiv. https://doi.org/10.48550/arXiv.2307.08823

Meta. (2025). *Frontier AI Framework*. https://ai.meta.com/static-resource/meta-frontier-ai-framework/?utm_source=newsroom&utm_medium=web&utm_content=Frontier_AI_Framework_PDF&utm_campaign=Our_Approach_to_Frontier_AI_blog

METR. (2025). Common Elements of Frontier AI Safety Policies. *METR Blog*.

https://metr.org/blog/2025-03-26-common-elements-of-frontier-ai-safety-policies/

Microsoft. (2025). *Frontier Governance Framework*. https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Microsoft-Frontier-Governance-Framework.pdf

NIST. (2024). *A Plan for Global Engagement on AI Standards*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-5.pdf

OpenAI. (2025). *Preparedness Framework*. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

Pistillo, M. (2025). *Towards Frontier Safety Policies Plus* (No. arXiv:2501.16500). arXiv. https://doi.org/10.48550/arXiv.2501.16500

Shevlane, T. (2022). *Structured access: An emerging paradigm for safe AI deployment* (No. arXiv:2201.05159). arXiv. https://doi.org/10.48550/arXiv.2201.05159

Stein, M., Bernardi, J., & Dunlop, C. (2024). *The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI* (No. arXiv:2410.04931). arXiv. https://doi.org/10.48550/arXiv.2410.04931

Stix, C., Pistillo, M., Sastry, G., Hobbhahn, M., Ortega, A., Balesni, M., Hallensleben, A., Goldowsky-Dill, N., & Sharkey, L. (2025). *AI Behind Closed Doors: A Primer on The Governance of Internal Deployment* (No. arXiv:2504.12170). arXiv. https://doi.org/10.48550/arXiv.2504.12170

Yates, J., & Murphy, C. N. (2019). *Engineering Rules: Global Standard Setting since 1880*. JHU Press.