Technical Report

FRONTIER MODEL FORUM

Risk Taxonomy and Thresholds for Frontier AI Frameworks

REPORT SERIES

Implementing Frontier AI Frameworks

DATE

June 18, 2025

About the Report

This is the latest in a **series of technical reports** on frontier AI safety frameworks that will examine how the frameworks can be used effectively across different organizational contexts. The series intends to provide detailed insight into key components of these frameworks, incorporating lessons from early adopters while acknowledging areas where best practices continue to evolve.

Executive Summary

Frontier AI frameworks outline methodologies for identifying, managing and mitigating the potential for large-scale risks to public safety and national security that stem from frontier AI development and deployment. This report examines the rationale for including only select risk domains within frontier AI frameworks, the current practices used to identify which risks warrant inclusion and how developers define thresholds to keep these risks within acceptable levels. Although the risks, threat models and thresholds in existing frontier AI frameworks vary, this report represents a survey of common risk taxonomies and thresholds.

Frontier AI frameworks are especially valuable for risks that share the following features: the development or deployment of the frontier AI model creates heightened marginal risk, such that it is not merely a similarly effective alternative to existing tools; the potential impacts to public safety and security could be severe and felt on a large scale; there is a credible pathway to extreme harm; and impacts could be instantaneous or potentially irreversible once triggered. These features explain why select risks warrant the unique governance measures included in frontier AI frameworks.

Most frontier AI frameworks describe structured exercises that developers conduct to identify extreme risks proactively. Threat modeling – adapted from cybersecurity and national security domains – is a process for systematically anticipating and identifying how various threat actors might leverage frontier AI to achieve harmful outcomes. Central to this process is mapping pathways from frontier AI models to severe outcomes, moving from broad catastrophic contexts to specific outcomes where frontier AI might provide utility to adversaries. Beyond addressing known large-scale risks, some

developers conduct research into unknown or emerging risks. Frontier developers also regularly engage with external stakeholders across academia, government and industry to discuss extreme risks.

Consensus is emerging around several domains that warrant inclusion in frontier AI frameworks. These include: Chemical, Biological, Radiological, and Nuclear (CBRN) threats, where AI could lower barriers to developing weapons of mass destruction; advanced cyber threats, where AI could lower barriers to conducting attacks against critical infrastructure; and advanced autonomous behavior threats, which are novel and for which assessment frameworks are evolving. Across these domains, frontier AI frameworks use thresholds to help determine when additional assessments or safeguards become necessary and when developers should pause or otherwise restrict development and/or deployment. Though exact implementation varies across organizations, frameworks typically operate with two distinct types of thresholds that work in sequence. First, “enabling capability thresholds” identify abilities that could potentially enable extreme harms if the model is deployed without additional safeguards. Second, “acceptable training or deployment thresholds” determine whether a model that has crossed an enabling capability threshold can be safely deployed or trained further after implementing safeguards. This "if-then" structure is particularly valuable for identifying risks that have not yet materialized, as it creates concrete triggering conditions for escalating safety and security measures.

When determining their thresholds, in particular acceptable training or deployment thresholds, developers navigate between two different approaches to baseline risk: static historical standards (risk thresholds established at a specific point in time) and marginal risk assessments (considering the additional risk their model adds to the ecosystem). While static approaches provide consistency and clear benchmarks, they could handicap development if other actors proceed with less stringent safeguards. By contrast, while dynamic approaches allow responsiveness to changing risk landscapes, they may facilitate gradual risk escalation through "risk creep." Balancing these considerations remains an active area of research and discussion. Resolving this baselining challenge remains an open question that may benefit from harmonized practices and approaches across industry.

This report outlines current approaches, but frameworks will evolve as capabilities advance. Frontier Model Forum (FMF) members are advancing work in key areas, including: improving threat modeling techniques; developing standardized methods for assessing capabilities; establishing empirical methods to validate mitigation strategies; and determining acceptable risk-benefit tradeoffs, which requires inclusive public discourse and diverse stakeholder input to represent collective priorities.

Section 1

Preparing for Possible Extreme Risks from Frontier AI

ROADMAP

1.1 Introduction and Scope

Frontier AI frameworks outline methodologies for identifying, managing and mitigating the potential for large-scale risks to public safety and national security that stem from frontier AI development and deployment. These frameworks typically focus on a narrow subset of risks that are severe and irreversible, and stipulate clear thresholds at which deployment or further development of a model may be paused until sufficient mitigations are in place.

This report examines how FMF members develop their frontier AI frameworks. In particular, it examines the rationale for focusing on specific extreme risks within frontier AI frameworks, the current practices used to identify which risks warrant inclusion and how developers define thresholds to keep these risks within acceptable levels. The report does not cover methods for evaluating specific risks or implementing safeguards, which are or will be addressed in separate technical reports.

1.2 Why Frontier AI Frameworks Emerged

Many frontier AI developers maintain comprehensive AI risk management policies that address a broad spectrum of risks. However, there are some risks for which conventional risk management approaches may not be adequate. Frontier AI frameworks are designed to manage these risks by providing an additional layer of scrutiny specifically for the most severe, large-scale risks posed by the most advanced AI models. Unlike traditional risk management, these frameworks must also address the unique challenge of preparing for capabilities and risks that have not yet emerged.

Frontier AI frameworks first emerged in late 2023, when Anthropic published its initial Responsible Scaling Policy, OpenAI published its Preparedness Framework (Beta), and the research organization METR published an initial primer on them. The concept of frontier AI frameworks has been refined extensively over the past two years, with leading experts in industry, academia and government contributing to their further development and more than a dozen leading frontier AI firms publishing frameworks of their own, after signing the Seoul Commitments. At the time of publishing, all FMF members have a frontier AI framework: Amazon's Frontier Model Safety Framework, Anthropic's Responsible Scaling Policy, Google's Frontier Safety Framework, Meta's Frontier AI Framework, Microsoft's Frontier Governance Framework, and OpenAI's Preparedness Framework.

The development of frontier AI frameworks has also drawn on practices in high-risk industries. Nuclear energy, finance, and national security sectors all distinguish between manageable risks and those requiring exceptional precautions. Governments also use various frameworks to define severe or large-scale risks, from national risk registers to emergency planning scenarios. For example, the UK's National Risk Register classifies as catastrophic an event that causes tens of billions of pounds of economic damage or more than 1,000 fatalities. Likewise, in the US, FEMA defines a catastrophic incident as one resulting in "extraordinary levels of mass casualties, damage, or disruption severely affecting the population, infrastructure, environment, economy, national morale, and/or government functions."

Stratifying risks by severity and scale also allows organizations to allocate resources proportionally to potential likelihood and impact. This tiered approach to governance means both comprehensive risk management and appropriate special handling for potentially catastrophic scenarios, balancing safety and security with continued innovation.

1.3 When Frontier AI Frameworks are Appropriate

Frontier AI frameworks are especially valuable for risks that share the following features:

1. **Heightened Marginal Risk**: The development or deployment of the frontier AI model creates heightened risk, not merely a similarly effective alternative to existing tools (e.g., search engines). For example, by democratizing dangerous capabilities (e.g., enabling low-skilled actors to build chemical weapons previously restricted to well-resourced states) or introducing entirely new risk categories.

2. **Severity and Scale**: The potential impacts to public safety and security could be severe and felt on a large scale. Although no standardized quantitative definition exists across all AI developers, frontier AI frameworks prioritize risks with catastrophic potential, meaning severe harm to many people or large-scale economic damage (i.e., tens of billions of dollars).

3. **Credibility**: There is a credible pathway to extreme harm. Importantly, a risk can be credible even if it's unlikely to occur, and the catastrophic potential of some risks justifies including low-probability scenarios. However, unlike established industries with historical precedents, developers currently rely on limited evidence when evaluating potential harm pathways. Developers must, therefore, balance accounting for realistic capability advancements against becoming overly speculative.

4. **Velocity and Irreversibility**: Impacts could be instantaneous or potentially irreversible once triggered. These risks typically lack adequate warning signs before manifesting fully, meaning developers may not have sufficient time to implement mitigating or reactive measures.

These four characteristics differentiate extreme risks from other important AI concerns and explain why they warrant the unique governance measures included in frontier AI frameworks. The potential irreversibility and significant scale of some of the potential harms necessitate preventative approaches, more forward-looking assessment methods, and more stringent responses—including, in some cases, either limiting deployment or further development until adequate safeguards are in place.

The risks that frontier AI frameworks address are generally acknowledged across jurisdictions. In some cases, there may be established international laws and norms that are relevant to their governance—which is not necessarily the case for other kinds of AI risks, where cultural factors might shape how they are perceived. This universality also makes these risks especially well-suited to the methods in safety engineering and security engineering frameworks that have inspired many frontier AI frameworks – technical risk assessments, defined risk thresholds, and mitigation effectiveness testing.

1.4 Distinctive Challenges of Frontier AI Risk Analysis

Frontier AI frameworks draw from academic literature and risk management practices in sectors like nuclear energy and aviation. However, developers must also take into account the following unique challenges of advanced AI systems:

1. **General Purpose Technology**: Unlike other industries (e.g., nuclear power), which have a clearly defined use case, frontier AI is a general-purpose technology.
   This requires anticipating and prioritizing risks across a far broader scope than traditional safety and security frameworks typically address.

2. **Dual-Use Risks and Benefits**: Many frontier AI capabilities that provide significant benefits are closely related to those that could enable severe risks (e.g., biological research capabilities that advance drug discovery could also lower barriers to developing biological weapons). This relationship between beneficial and potentially harmful applications adds complexity to risk management decisions, as developers must balance preserving valuable capabilities while mitigating potential misuse.

3. **Forward-Looking Assessment**: Risk assessment for extreme risks from frontier AI is typically designed in anticipation of future, more capable systems, under the assumption that AI capabilities will continue to advance in a wide range of domains. This forward-looking nature means frameworks rely more heavily on methodologies like threat modeling and expert consultation when assessing possible hazards and threats.

4. **Nascent Risk Quantification**: While risk estimates in more established domains (e.g., nuclear engineering) are typically defined in probabilistic, numerical terms – often based on historical base rates of component failure – frontier AI risk estimates are more qualitative, reflecting a larger degree of uncertainty. In cases where they are quantitative, these numbers are often estimates based on expert consultation rather than historical data.

5. **Adversarial Dynamics**: Frontier AI frameworks tend to have a greater focus on risks that arise from misuse than is common in other sectors. In industries such as aviation, risk management frameworks are more tailored towards product safety and prevention of unintentional failures, rather than intentional misuse.

Frontier AI frameworks also accommodate these sources of uncertainty through iterative design, allowing risks to be added or amended as understanding evolves. As demonstrated by Anthropic, Google, and OpenAI in their second-generation frameworks, these updates may include revising in-scope risks based on an evolving understanding of frontier AI risks.

Section 2

Methodologies for Identifying Extreme AI Risk

Frontier AI frameworks address severe or extreme risks, with some distinguishing between intentional misuse risks and inadvertent risks arising from model capabilities. Most frontier AI frameworks describe structured exercises that developers conduct to identify extreme risks proactively.

2.1 Threat Modeling

Threat modeling – adapted from cybersecurity and national security domains – is a process for systematically anticipating and identifying how various threat actors might leverage frontier AI to achieve harmful outcomes and mapping the potential pathways to those outcomes.

At the time of writing, threat modeling for frontier AI risks involves a distributed effort across developers, independent researchers, and domain experts. This work takes many forms: developers conduct in-house research, commission studies from external experts, collaborate with academic institutions, and draw on findings from independent researchers.

The process typically begins with preliminary investigations and expert consultations to identify potential severe risks worth deeper examination. Analysts may start with contexts having catastrophic potential (such as the deployment of a biological weapon or a cyber attack causing a financial market crash) and examine how frontier AI might enable them, or begin with frontier model capabilities and project how they might be misused or inadvertently lead to harm. Effective threat modeling typically combines both approaches.

Once promising areas for investigation are identified, researchers conduct deeper dives into specific risk domains. These investigations employ various methodological tools, including:

- Forecasting studies to project capability development timelines and potential impacts
- Expert consultations and surveys to gather domain-specific knowledge
- Workshops bringing together AI researchers with experts from relevant fields (e.g., biosecurity, cybersecurity)
- Tabletop exercises to simulate how risks might unfold in practice
- Historical analysis of analogous technologies and their impacts

Central to this process is mapping pathways from frontier AI models to severe outcomes. This refinement moves from broad catastrophic contexts to specific outcomes where frontier AI might provide utility to adversaries. For example, this might include the ability to assist in the development of known chemical weapons amongst low and moderate-skilled actors, or the development of novel biological weapons amongst high-skilled actors. Through this distributed approach, the community develops detailed threat scenarios that specify how adversaries might use frontier AI to achieve severe outcomes, identifying specific tasks, model capabilities to exploit, and complementary tools. These scenarios directly inform both risk assessment methodologies and protective mitigations against various adversary types.

2.2 Monitoring for Emerging Risks

Beyond addressing known large-scale risks, some developers conduct research into unknown or emerging risks. This might take the form of investigating how models are used once deployed, or consultation with external experts from domains beyond those currently in scope. This broad monitoring approach helps uncover less obvious sources of extreme risk. The four features outlined earlier provide a useful means for analyzing newly identified risks and determining if they warrant formal inclusion.

2.3 Information Sharing

Frontier developers regularly engage with external stakeholders across academia, government, and industry to discuss extreme risks. Collaboration mechanisms include:

- Forums for voluntary responsible disclosure between frontier developers
- Academic conferences and workshops
- Partnerships between developers and external experts

Voluntary information exchange is valuable for advancing collective risk identification and mitigation capabilities.

2.4 Current Domains of Consensus

Consensus is emerging around the domains that warrant inclusion in frontier AI frameworks. As understanding of frontier AI risks evolves, additional domains may reach a similar consensus status, and existing domains may be amended to reflect the latest understanding. The iterative nature of frontier AI frameworks allows for such changes as evidence and expert consensus develops. The frameworks of FMF members all include risks in two or more of the following domains:

1. **Chemical, Biological, Radiological, and Nuclear (CBRN) Threats**: Frontier AI could potentially lower barriers to accessing or developing weapons of mass destruction through capabilities like providing specialized technical knowledge, identifying novel agents, or optimizing production methods. CBRN scenarios are classified as extreme risks because they could enable threat actors with limited resources to cause severe, widespread, and irreversible harm. Developers focus particularly on scenarios where AI could enable less-resourced actors to develop capabilities previously limited to a small number of well-resourced actors.

2. **Advanced Cyber Threats**: Frontier AI could potentially enhance the capabilities of threat actors in identifying vulnerabilities, developing exploits, or executing sophisticated attacks against critical infrastructure or systems. Cyber attacks facilitated by frontier AI could affect multiple critical systems simultaneously, potentially leading to widespread disruption to essential services with significant economic and societal consequences. Framework evaluations typically assess whether models could enable novel attacks that bypass current defenses or dramatically reduce the expertise required for sophisticated operations.

3. **Advanced Autonomous Behavior Threats**: Unlike CBRN and cyber threats with historical precedents, autonomous behavior risks are entirely novel and lack established assessment frameworks. This could include AI systems using their capabilities in ways that conflict with human intentions or values, potentially pursuing objectives that emerge during training but weren't explicitly programmed; or systems improving themselves rapidly (e.g., fully automating the AI R&D pipeline), potentially compressing timelines for addressing other risks and outpacing governance mechanisms.

All consensus is emerging around inclusion of the domains above within frontier AI frameworks, specific approaches to risk manifestation, likelihood assessment, and mitigation strategies vary across organizations to accommodate different development and deployment contexts.

Section 3

Setting Thresholds for Extreme AI Risks

ROADMAP

Frontier AI frameworks use thresholds to help determine when additional assessments or safeguards become necessary, and when development and/or deployment should be restricted.

3.1 The Role and Purpose of Thresholds

1. **Enabling Capability Thresholds** (also called "critical capability levels" or "capability thresholds"): Abilities that could potentially enable extreme harms if the model is deployed without additional safeguards. Meta defines these as capabilities "essential to enabling the realization of a threat scenario," such as "PhD level proficiency in biology," in a scenario that could facilitate the development of biological weapons. Other examples include Amazon's "Critical Capability Thresholds," Google's "Critical Capability Levels," Anthropic's "Capability Thresholds," OpenAI's "Capability Thresholds," and Microsoft's "Capability Thresholds." While the presence of these capabilities doesn't guarantee that harmful outcomes will occur in the planned deployment setting, it signals entry into a new phase of heightened risk, where more rigorous risk assessments for this domain and stronger baseline safety and security measures are potentially warranted.

2. **Acceptable Development or Deployment Thresholds** (also called "outcome-focused thresholds" or "residual risk thresholds"): Criteria that determine whether a model that has crossed an enabling capability threshold can be safely deployed or trained further after implementing safeguards. Meta's outcomes-focused approach assigns models to "Moderate," "High," or "Critical" risk thresholds based on their potential to enable catastrophic scenarios in the proposed deployment context (i.e., closed deployment, limited release, or open weights release). OpenAI assesses whether

"safeguards sufficiently minimize the associated risk," and Anthropic requires a "Safeguards Assessment" to verify risks have been adequately mitigated. Microsoft evaluates whether "risk mitigation measures effectively reduce potential harms" through pre-deployment testing. Amazon evaluates whether safeguards "appropriately mitigate the risks (e.g., by preventing reliable elicitation of the capability by malicious actors)." *Judgments about whether a particular deployment or training run meets the acceptable training or deployment thresholds integrate capability assessments, expected safeguard effectiveness, and societal resilience.*

If the enabling capability thresholds above are reached, developers may introduce more extensive consideration of acceptable training or deployment thresholds and related risk governance procedures. This "if-then" structure is particularly valuable for identifying risks that have not yet materialized, as it creates concrete triggering conditions for escalating safety and security measures, avoiding both premature investment in safeguards and delays in implementing appropriate protections when capabilities begin to approach critical thresholds.

3.2 Establishing Risk Baselines

When determining thresholds, in particular acceptable training or deployment thresholds, developers navigate between two different approaches to baseline risk:

- **Static Historical Standards**: Under this approach, risk thresholds would be established at a specific point in time (e.g. 2023) and never adjusted, regardless of changes in the ecosystem. While this provides consistency and clear benchmarks, it could handicap development if other actors proceed with less stringent safeguards, and be unsustainable for developers to maintain in the absence of industry-wide or international standards around risk. It also fails to acknowledge that real-world threat landscapes naturally evolve over time, in some cases unrelated to AI development and deployment.

- **Marginal Risk Assessments**: On the other hand, developers would consider the additional risk their model adds beyond what's already possible in the ecosystem. While dynamic approaches allow responsiveness to changing risk landscapes, they may facilitate gradual risk escalation through "risk creep" – where multiple models, each introducing only small marginal increases which do not cross individual "net new" thresholds, may collectively create significant risk growth over time as these small increases compound. Additionally, irresponsible actors might release risky models without consideration for ecosystem impact, shifting baselines for all actors.

Balancing these considerations remains an active area of research and discussion. Currently, developers have adopted varied approaches:

- OpenAI's Preparedness Framework explicitly incorporates marginal risk assessment, allowing for adjustment of safeguard requirements if another company has verifiably released systems with similar capabilities without comparable protections. However, this adjustment is only permitted if they can "rigorously confirm that doing so does not meaningfully increase the overall risk of severe harm," publicly acknowledge any adjustment, and continue to maintain more protective safeguards than competitors.

- Google's framework and security recommendations include explicit provisions that "our adoption of the protocols described in this Framework may depend on whether such organizations across the field adopt similar protocols." They emphasize that certain mitigations "should be understood as recommendations for the industry collectively" and that their "social value is significantly reduced if not broadly applied." Additionally, they may adjust security levels if "a model does not possess capabilities meaningfully different from other widely available models that have not demonstrably caused or contributed to severe risks."

- When conducting initial pre-mitigation assessments of risk, Anthropic establishes a fixed 2023 baseline. When considering post-mitigation risk, the framework outlines a default set of required safeguards that are expected to bring risk down to acceptable levels, while also including provisions for reconsidering safeguard requirements "if another actor in the frontier AI ecosystem will pass, or be on track to imminently pass, a Capability Threshold without implementing measures equivalent to the Required Safeguards." In such cases, they commit to maintaining "more conservative safeguards than the other AI developer" and to advocating for regulatory action.

- Microsoft's Frontier Governance Framework considers "the marginal capability uplift" a model may provide over and above currently available tools and information, including currently available open-weights models.

- Meta's approach emphasizes "net new" outcomes – focusing on whether capabilities enable outcomes not previously achievable with existing tools and resources, including other frontier models.

- Amazon's Frontier Model Safety Framework assesses whether a model provides a "material 'uplift' in excess of other publicly available research or existing tools."

The iterative nature of frameworks also allows for periodic threshold review and updates, accommodating changes to ecosystem risk and broader evolutions in societal tolerance for risk from frontier AI models. Resolving this baselining challenge remains an open question that would benefit from harmonized practices and approaches across industry. One approach is for developers to commit to maintaining safety standards as stringent as the current ecosystem baseline, combined with sharing sufficient information, either publicly or privately, about implemented safeguards to enable such decisions.

Section 4

Continuing Work

REQUEST FOR COMMENT

We welcome engagement with this and forthcoming technical reports from across the frontier AI safety and security ecosystem. Researchers and organizations interested in further refining and harmonizing the implementation of safety frameworks are invited to reach out to the Frontier Model Forum.

Please offer feedback at: info@frontiermodelforum.org

This report outlines current developer approaches, but as capabilities advance and our risk management abilities improve, frameworks will be updated accordingly. Frontier AI developers are advancing work in key areas, including:

1. **Improving Risk Identification Methodologies**. This includes improving threat modeling techniques and establishing more robust methodologies for assessing the potential realization of different threat scenarios.

2. **Advancing Capability Evaluation and Mitigation Efficacy.** This includes developing standardized methods for assessing capabilities and risk, establishing empirical methods to validate mitigation strategies, and building shared tools that can be used across the industry. This work will be explored further in forthcoming FMF technical reports.

3. **Balancing Risk Tolerance with Potential Benefits**. What level of frontier AI risk is acceptable? How should potential benefits be weighed against possible harms? These questions extend beyond technical considerations to encompass societal values, ethical frameworks, and democratic decision-making. While developers can provide transparency about their thresholds and decision processes, the ultimate determination of acceptable risk-benefit tradeoffs requires inclusive public discourse, diverse stakeholder input, and appropriate governance structures to represent collective values and priorities.

Future research should focus on developing evidence-based approaches for continuous calibration of risk management strategies, ensuring they remain responsive to both empirical findings and evolving societal expectations.