ISSUE BRIEF

FRONTIER MODEL FORUM

Issue Brief: Frontier AI Biosafety Thresholds

DATE

May 12, 2025

The rapid advancement of frontier AI presents transformative opportunities in the biological domain. For example, frontier AI systems may accelerate the discovery of new medical treatments, optimize biomanufacturing processes, or facilitate the development of novel biocatalysts. However, given that frontier AI capabilities are often dual-use, they may also heighten risks from misuse in the biological domain, such as by lowering barriers to creating known or even novel biological threats. As frontier AI grows increasingly capable, it is crucial to develop risk management practices that enable society to harness the benefits of frontier AI while proactively managing its most significant risks.

Frontier AI safety frameworks, which establish thresholds for when a model's capabilities may require additional scrutiny and/or safeguards, have become an essential tool for developers seeking to manage severe risks responsibly. Yet how to set and evaluate thresholds for biosafety risks can be challenging. Based on expert discussions among Frontier Model Forum (FMF) member firms and the wider biosafety community, this issue brief highlights emerging industry consensus on a core biosafety threshold for frontier AI. In addition to specifying the point at which an AI model's capabilities in biological domains may require further assessments and/or enhanced safeguards, it also provides initial guidance for evidence-based threshold determinations.

CURRENT CONSENSUS ON BIOSAFETY THRESHOLDS

Frontier AI thresholds describe predefined notions of risk that indicate when additional action is warranted to avoid unacceptable outcomes. As noted in a separate issue brief, there are several potential ways to establish such thresholds. *Capability thresholds* identify specific AI capabilities relevant to a threat scenario which, absent effective mitigation measures, may pose unacceptable levels of risk to society. *Risk thresholds* specify a level of unacceptable risk in quantitative terms, such as the likelihood or severity of an outcome. *Compute thresholds* specify a level of processing power used to train a model or system and can serve as an ex ante proxy for capabilities. Finally, *outcome-based thresholds* define threat scenarios describing how a malicious user might use frontier AI to realise an unacceptable outcome.

Significantly, **capability thresholds have emerged as the most commonly used type of threshold for biosafety in frontier AI frameworks.** Although capability thresholds are a less direct measure of risk than risk thresholds, they are a better proxy for risk than compute thresholds and more straightforward to measure than risk thresholds.[1] As a result, capability thresholds offer a compelling compromise that makes them well-suited for establishing biosafety thresholds. Similarly, outcome-based thresholds also offer a way of linking models with ultimate harms. By testing the extent to which access to a frontier AI model would enable a malicious actor to achieve predefined threat scenarios, developers can identify relevant incremental thresholds (i.e. critical, high, medium) proportional to the utility the model provides towards realizing that outcome that serve as effective proxies of risk.[2]

What is most notable about the biosafety thresholds in the frontier AI frameworks published to date is that they typically include slight variations on the same core capability or outcome-based threshold, namely: **whether an AI model or system is capable of significantly enabling human individuals or groups with limited scientific expertise to obtain, produce, or deploy biological threats, compared to other available resources.**[3]

Threshold Type	Description	Trade-off
Capability threshold	Identifies specific AI capabilities relevant to realizing a bio-related threat scenario which, absent effective mitigation measures, may pose unacceptable levels of risk to society	More measurable than risk thresholds and a better risk proxy than compute thresholds, but a less direct measure of ultimate risk compared to risk thresholds
Compute threshold	Sets a maximum level of computational power (e.g., FLOPs) used for training a model as an ex ante proxy for bio-related capabilities	Relatively straightforward to measure, but an imperfect proxy for actual risk, as greater compute doesn't always equate to greater risk
Outcome-based threshold	Defines specific bio-related threat scenarios (e.g., AI assisting a non-expert in creating a bioweapon) and assesses the AI's potential contribution	Connects AI capabilities directly to potential real-world harms, but comprehensive and realistic scenarios can be challenging to define and evaluate
Risk threshold	Defines an unacceptable level of risk quantitatively, based on the estimated likelihood and severity of a bio-related harm	Directly measures risk, but difficult to implement reliably due to the complexity and uncertainty in estimating future risks

Although some frontier AI safety frameworks include additional thresholds as well,[4] the convergence around a "non-expert uplift" threshold for biosafety reflects an emerging consensus on biosafety thresholds.[5] Among frontier AI firms and the broader AI biosafety community, AI models are thought to cross a critical biosafety threshold when they provide assistance in the form of specialized knowledge, troubleshooting guidance, or procedural instruction that meaningfully reduces the expertise, resources, or time traditionally required for low-skill threat actors to create biological threats.[6] However, informational uplift alone may not be sufficient to overcome the necessary practical requirements for biological threat creation, like obtaining access to wet lab facilities or acquiring dangerous materials. As a result, the consensus biosafety threshold above should primarily be used to inform decisions about conducting further risk assessments or implementing further safeguards, and should be considered in conjunction with other established biosafety thresholds.

KEY CONSIDERATIONS FOR FRONTIER AI BIOSAFETY THRESHOLDS

As noted above, AI models or systems capable of significantly enabling non-expert actors to create biological threats, as compared to other baseline resources, constitutes an important risk scenario for further analysis.

This consensus threshold highlights several key components for consideration:

- **Threat Actor Uplift**. All current thresholds focus on frontier AI's ability to provide assistance or "uplift" that enhances human capabilities beyond what would otherwise be possible, bridging the expertise gap between individuals or groups with limited scientific training and those with specialized knowledge necessary for creating biological threats.
- **"Significant" Uplift**. All thresholds qualify the level of uplift as "significant," "meaningful," or "material." The ambiguity in defining "significant" means that in practice, developers must still exercise judgment regarding what level of uplift is either deemed acceptable or would trigger further analysis or risk mitigation measures. Further, the inherent subjectivity may complicate efforts to establish consistent and reliably comparable safety benchmarks across organizations. Moving forward, it may be useful to establish consensus on what constitutes "significant" uplift.

- **Comparative Baselines**. Some current thresholds assess the increase in risk by considering the difference between what threat actors could achieve with AI assistance compared to other existing resources, with some using online resources available in a baseline year[7] or other publicly available research or tools currently available[8] as comparisons. Setting a comparative baseline is necessary to measure marginal risk at the time of measurement. Using a specific baseline year (e.g., 2023-level resources) can be useful to prevent a "shifting goalpost" scenario where, as general resources improve over time, the perceived relative impact of new AI models might stay constant or decrease, while potentially masking dangerous increases in their absolute capabilities that may suggest a need for a societal response. However, given that ecosystem risk may evolve over time, this approach may be too inflexible to remain robust in the long term. Determining an appropriate baseline for measurement remains an open question in biosafety thresholds.

- **Threat Actor Expertise**. Some thresholds differentiate between AI's impact on different types of actors (e.g., no scientific background, general STEM background, or advanced biology degrees), recognizing that different levels of assistance may be concerning for different threat actor profiles.[9] Currently, there is only full consensus on the importance of measuring uplift for low-skilled actors developing bioweapons. However, there may be other important thresholds to monitor as well. Several firm biosafety thresholds also note the importance of determining when frontier AI can provide meaningful uplift to high-skilled actors (Ph.D.-level education in relevant domain), potentially housed in small groups or moderately-resourced state programs, to develop and deliver a *novel* or *highly dangerous* biological weapon.[10] For example, an AI model that helps a trained biologist design gene-editing experiments more efficiently may cross a critical threshold if it enables them to create a highly dangerous, novel pathogen more effectively than would otherwise be possible.

- **Real-World Feasibility**. Some biosafety thresholds account for whether AI assistance enables "reliable" or "consistent" execution of complex biological protocols, adding real-world operational feasibility to theoretical knowledge uplift.[11] This may be an important feature of measuring the plausibility of real-world harm resulting from frontier AI capabilities. Real-world feasibility should also take into account the holistic supply chain for end-to-end threat production, including existing downstream mitigations and protections that may exist at various points, many of which are unrelated to AI development.

The above considerations are all relevant for establishing a "non-expert uplift" threshold within frontier AI safety frameworks. However, it is important to note that several frameworks employ multiple, tiered thresholds (e.g., indicating medium, high, or critical risk levels) that trigger corresponding mitigation actions well before intolerable risk levels are reached. Future work may explore other threat models and their associated thresholds in more detail.

PRELIMINARY CONSENSUS ON EVIDENCE FOR THRESHOLD DECISIONS

Frontier capability assessments are procedures conducted on frontier AI models to gather evidence of whether they have capabilities that could increase risks to public safety and security. For bio-related risks, biosafety evaluations may be run as part of a capability assessment approach designed to produce evidence indicating whether a model could assist in developing a biological weapon.

Current best practice suggests threshold determinations for biosafety should be made on the basis of *cumulative* evaluation evidence, structured in a holistic assessment approach. Since the results from a single evaluation are unlikely to indicate unequivocally whether a biosafety threshold has been crossed, threshold determinations should be made on the basis of multiple biosafety evaluations and sources of evidence. The field is still nascent enough that it remains unclear which precise combination of evidence is needed to determine whether a threshold has been reached or passed.

However, several capabilities may serve as strong indicators of whether the core "non-expert uplift" biosafety threshold may have been reached or passed. Given existing bottlenecks to biological threat creation, it is helpful to carry out bottleneck assessments designed to provide insight into whether an AI model or system can perform the following tasks significantly better than other available baseline resources:

- Generate or assess ideas for known or novel biological threat creation by carrying out expert-level scientific reasoning (e.g., generating hypotheses deemed highly plausible and predicting the results of unpublished experiments)
- Assist with the acquisition of materials and equipment required for the creation of a known or novel biological threat
- Assist with the design and formulation of biological weapons (e.g., by troubleshooting protocols)
- Assist with the release or deployment of biological weapons
- Carry out autonomous biological R&D (e.g., independently designing, planning, and potentially executing research cycles to create known biological compounds or organisms)

These capabilities should not be viewed in isolation, but rather as a constellation of factors that together may indicate whether a model crosses critical biosafety thresholds. Developers and deployers should consider how these capabilities interact and potentially amplify risk when deployed in real-world contexts, particularly when accessible to users with varying levels of expertise and intent. For more on how to evaluate models for the capabilities above, see our preliminary taxonomy of biosafety evaluations or Appendix B below.

CONCLUSION

Frontier AI safety frameworks establish thresholds that specify when additional risk assessments or mitigation measures should be implemented during AI development. For bio-related risks, frontier AI frameworks typically include a capability or outcome-based threshold focused on whether a model significantly enables non-experts to create biological threats compared to other available resources.

As frontier AI continues to advance, establishing, refining, and assessing these thresholds will become increasingly important. Many open questions remain, including how to precisely define "significantly enable" in the context of biosafety, the cumulative amount of evidence needed to determine when a threshold has been crossed, and determining how model performance on assessments translates to risks. Further research and cross-industry collaboration, in particular with domain experts, are needed to address these questions and enable frontier AI developers to implement biosafety framework thresholds more effectively.

APPENDIX A: FMF MEMBER FIRM BIO THRESHOLDS

Note: This table only highlights thresholds where there is key consensus. We omit the "low" risk level that some frameworks include. Thresholds are accurate as of May 2025.

Company	Threshold
Amazon	AI at this level will be capable of providing expert-level, interactive instruction that provides material uplift (beyond other publicly available research or tools) that would enable a non-subject matter expert to reliably produce and deploy a CBRN weapon.
Anthropic	CBRN-3: The ability to significantly assist individuals or groups with basic STEM backgrounds in obtaining, producing, or deploying CBRN weapons. Assessed by comparing what potential attackers could achieve with full model access versus 2023-level online resources, assuming they have funding and up to one year of time to invest, but no initial specialized expertise.
	CBRN-4: The ability to substantially uplift CBRN development capabilities of moderately resourced state programs (with relevant expert teams), such as by novel weapons design, substantially accelerating existing processes, or dramatic reduction in technical barriers.
Google DeepMind	Can be used to significantly assist a low-resourced actor with dual-use scientific protocols, resulting in a substantial increase in ability to cause a mass casualty event.
Meta[12]	Small cells of low or moderate skill actors acquire sufficient scientific and operational capabilities to succeed at complex chemical and biological workflows.
	Small cells of low and moderate skill actors can consistently execute relevant and complex wet-lab biological protocols.

Microsoft	Medium: The model provides some assistance increasing the ability of an existing expert (Ph.D.-level education in related fields) to develop a known CBRN threat.
	High: The model provides a meaningful uplift to an expert's (Ph.D.-level education in related fields) ability to develop and deliver a novel CBRN threat, OR the model provides a meaningful uplift to a medium-skilled actor's (e.g., STEM education) ability to develop and deliver a known CBRN threat.
	Critical: The model provides a meaningful uplift to an expert's ability to develop a highly dangerous novel threat or significantly lowers the barriers to a low-skilled actor developing and delivering a known CBRN threat**.**
OpenAI	High: The model can provide meaningful counterfactual assistance (relative to unlimited access to baseline of tools available in 2021) to "novice" actors (anyone with a basic relevant technical background) that enables them to create known biological or chemical threats.
	Critical: The model can enable an expert to develop a highly dangerous novel threat vector (e.g., comparable to a novel CDC Class A biological agent) OR the model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel biological threat without human intervention.

APPENDIX B: EVALUATION TECHNIQUES FOR BIOSAFETY CAPABILITIES

Certain evaluation techniques may be particularly well-suited to gathering evidence about specific biosafety-relevant capabilities. For example:

- **Benchmark evaluations and representative lab tasks** are useful for demonstrating whether an AI model or system can carry out biological R&D processes autonomously, possess potentially harmful biological knowledge, or facilitate workflows relevant to biological weapons production.
- **Red-teaming exercises with biosafety experts** are well-suited to show whether an AI system is able to generate or assess ideas for biological threat creation. Uplift studies (which do not require biosafety experts) may also be useful to elicit this evidence.
- **Uplift studies** are well-suited to showing whether an AI system is able to assist non-experts in troubleshooting protocols related to the production of bioweapons. They may also provide insight into whether an AI system can help with the acquisition of materials and equipment required for the creation of a biological threat, support the release or deployment of biological weapons, or facilitate operational execution of the attack and avoid law enforcement detection.

FOOTNOTES

1. For example, a biosafety capability threshold might be set at the point where an AI model can autonomously design novel synthetic pathogens with pandemic potential, whereas a biosafety compute threshold would only measure processing power used to train a frontier AI model, regardless of whether the model could perform specific biosafety-relevant tasks.
2. See Meta's Frontier AI Framework, pg. 15.
3. These include: 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. For convenience, these are also listed below in Appendix A.
4. See Anthropic, Meta, Microsoft, and OpenAI's frameworks.
5. Expert discussions involving academia, industry, and government participants were held to discuss biosafety thresholds.
6. For similar discussions, see also Sandbrink, 2023; Bloomfield et al., 2024; and RAND, 2024.
7. Anthropic pg. 4 and OpenAI pg. 5.
8. Amazon pg. 2 and Meta pg. 15.
9. Meta, Microsoft, and OpenAI.
10. Meta, Microsoft, OpenAI.
11. Amazon and Meta.
12. Note that several threat scenarios were omitted.