ISSUE BRIEF

FRONTIER MODEL FORUM

Thresholds for Frontier AI Safety Frameworks

DATE
February 7, 2025

Although frontier AI holds enormous promise for society, advanced AI systems may also pose significant risks to national security and public safety. Frontier AI safety frameworks have recently emerged as a method for frontier AI developers to demonstrate how they manage those risks effectively. By establishing processes for how to identify, evaluate, and mitigate severe risks, safety frameworks offer a principled approach for the responsible development and deployment of frontier AI, in a way that keeps risks within tolerable levels.[1]

Thresholds are an essential element of safety frameworks. Regardless of their overall approach and structure, all safety frameworks establish predefined thresholds that indicate when the potential risks of a given model or system warrant deeper inspection–and in some cases, heightened safeguards–to avoid unacceptable outcomes. Thresholds thus inform key decisions about further model development and deployment. Establishing thresholds for safety frameworks is a challenging task given the complex and fast moving nature of advanced AI, the nascent understanding of AI risk, and the numerous factors that contribute to these risks.

This issue brief seeks to advance and inform public understanding of frontier AI thresholds. Drawing on insights from experts within the FMF as well as the broader AI safety and security community, this brief elaborates on the importance of thresholds for frontier AI safety frameworks and outlines the different types of thresholds that have been proposed.

UNDERSTANDING FRONTIER AI THRESHOLDS

Within AI safety frameworks, thresholds describe predefined notions of risk that indicate when additional action is warranted to avoid unacceptable outcomes.[2] Although they may be operationalized in different ways, thresholds provide a structured approach to managing risk by establishing clear boundaries for acceptable levels of risk and a process to keep risks within tolerable levels.[3] Setting thresholds in advance of development enables all actors across the frontier AI development lifecycle, including developers, deployers, and regulators, to play their part in addressing any potential hazards before they materialize. This is especially important given the types of risks that safety frameworks are typically designed for, including Chemical, Biological, Radiological, and Nuclear (CBRN) and advanced cyber risks that pose severe harms and may take significant resources and lead-time to address in advance of deployment.

The process of setting thresholds also helps to create accountability for all actors in the AI development lifecycle. By establishing concrete standards for acceptable levels of risk and identifying key outcomes of concern, thresholds enable more efficient internal decision-making processes by frontier AI developers and providers, as well as any potential external oversight that may be required. Without predefined thresholds, decisions about AI safety measures may become more ad hoc, making it difficult to weigh tradeoffs around the benefits and risks associated with frontier AI models and systems.

TYPES OF FRONTIER AI THRESHOLDS

There are several main approaches to establishing thresholds within frontier AI safety frameworks, each of which entails a unique set of tradeoffs. These approaches include:

- ***Compute thresholds*** are defined in terms of the computational resources used to train a model. Compute may be considered a proxy for risk because, to date, increases in the amount of computational resources used to train frontier AI models have been correlated with advances in model capabilities.[4] This suggests that models trained with larger amounts of compute may introduce a higher risk of intolerable outcomes stemming from increased model capabilities.

  That said, while compute thresholds are relatively straightforward to understand and measure, they are an imperfect proxy for risk. Recent algorithmic progress has demonstrated that it may be possible to create high-risk systems with less compute than previously believed. Focusing on only models above a certain compute threshold may exclude smaller models that could possess potentially harmful capabilities, and conversely may inundate evaluators with larger models that have only benign capabilities.[5] Compute thresholds are therefore at best used as an initial filter for identifying models that may warrant further scrutiny.[6]

- ***Risk thresholds*** set explicit limits for acceptable levels of the estimated risk stemming from the deployment of a frontier AI model or system. They are typically expressed through both the *likelihood* (as a probability estimate) and the *magnitude* or degree of harm (e.g., fatalities, economic damage).[7] While risk thresholds provide the most direct link between risk and societal impact, they are currently challenging to implement.

  Due to the multi-faceted and inherent dual-use nature of many large-scale frontier AI models and their resulting societal effects, producing reliable risk estimates is a complex and imprecise task. Risk thresholds have been successfully implemented in other industries, such as aviation, where regulatory bodies set specific acceptable levels of risk (e.g., probability of catastrophic failure per flight hour) based on empirical data and well-understood failure modes. Setting these thresholds for frontier AI is significantly more challenging due to the lack of historical data, the potential for novel and unprecedented failure modes, and the difficulty in modeling complex socio-technical interactions. Moreover, setting risk thresholds will require engaging in difficult discussions around normative tradeoffs and should follow a multilateral discussion involving stakeholders from across the AI lifecycle. However, if and when robust estimation methods are available, risk thresholds have the promise to offer a more direct foundation for decision-making.

- ***Capability thresholds*** identify specific capabilities at which, absent mitigation measures, models or systems may pose unacceptable levels of risk to society. For example, an AI system that is able to provide clear instructions about how to synthesize highly lethal and transmissible pathogens may pose an unacceptable level of risk to society. Capability thresholds provide a more direct link to potential hazards than compute measurements and are currently easier to measure than quantitative risk assessments. That said, using capability thresholds alone may miss important contextual factors about how systems could be used post-deployment that would change their societal impact.

While the thresholds above have been the most prominently discussed to date, other approaches continue to emerge. For instance, ***outcome-based thresholds*** specify a set of outcomes that represent the intolerable threshold, along with a set of threat scenarios that describe how a frontier AI model could be misused to produce those outcomes.[8] Assessments are designed to test whether a frontier AI model could uniquely enable a threat scenario, and to measure the level of risk a model poses towards realizing these intolerable outcomes. This approach seeks to accommodate the dual-use nature of frontier AI capabilities, by focusing on durable scenarios that an AI developer can evaluate against and seek to avoid or mitigate.

Thus far, capability thresholds have emerged as the most commonly used type of threshold in frontier AI safety frameworks. Although capabilities are an indirect proxy for risk, capability thresholds can be more directly linked to risk than compute thresholds. Moreover, capability thresholds are more straightforward to measure than risk thresholds, which are currently difficult to estimate for frontier AI. Capability thresholds therefore offer an effective compromise between risk and compute thresholds, and can inform determinations about the safety of a given AI system even when risk estimates remain uncertain. Outcomes-based threshold may also serve as a promising compromise between risk and compute thresholds, as well as a conceptual link between capability and risk thresholds.

CONCLUSION

Thresholds are a critical component of frontier AI safety frameworks, providing clear guidelines for when additional safety measures must be implemented during AI development. Compute, risk, outcome, and capability thresholds each offer distinct advantages, balancing different approaches to measuring and managing risks.

The effective implementation of critical thresholds will require a concerted effort. As frontier AI continues to advance, establishing, refining, and assessing these thresholds will become increasingly important. Many open questions still remain with respect to thresholds, including how to make determinations about when thresholds are crossed and how different thresholds may be used in tandem. Further research is needed to address these questions and enable frontier AI developers and providers to implement safety framework thresholds more effectively.

FOOTNOTES

[1] For more on these risks, see the Frontier AI Safety Commitments announced at the AI Seoul Summit in May 2024.

[2] See Outcome 1.II of the Frontier AI Safety Commitments for more on how thresholds can be defined.

[3] Koessler et al. 2024, pg. 6

[4] See for example Google DeepMind's LLM scaling analysis.

[5] Hooker 2024

[6] Heim & Koessler 2024

[7] Koessler et al. 2024

[8] See Meta, "Our Approach to Frontier AI." February 3, 2025.