METR

Common Elements of Frontier AI Safety Policies

December 2025

Summary

A number of developers of large foundation models have committed to corporate protocols that lay out how they will evaluate their models for severe risks and mitigate these risks with information security measures, deployment safeguards, and accountability practices. Beginning in September of 2023, several AI companies began to voluntarily publish these protocols. In May of 2024, sixteen companies agreed to do so as part of the Frontier AI Safety Commitments at the AI Seoul Summit, with an additional four companies joining since then. Currently, twelve companies have published frontier AI safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI, and NVIDIA.

This December 2025 version contains references to some developers' updated frontier AI safety policies. In addition, it also mentions relevant guidance from the EU AI Act's General-Purpose AI Code of Practice and California's Senate Bill 53.[1]

In our original report from August 2024, we noted how each policy studied made use of *capability thresholds*, such as the potential for AI models to facilitate biological weapons development or cyberattacks, or to engage in autonomous replication or automated AI research and development. The policies also outline commitments to conduct model evaluations assessing whether models are approaching capability thresholds that could enable severe or catastrophic harm.

When these capability thresholds are approached, the policies also prescribe *model weight security* and *model deployment mitigations* in response. For models with more concerning capabilities, we noted that in each policy developers commit to securing model weights in order to prevent theft by increasingly sophisticated adversaries. Additionally, they commit to implementing deployment safety measures that would significantly reduce the risk of dangerous AI capabilities being misused and causing serious harm.

The policies also establish *conditions for halting development and deployment* if the developer's mitigations are insufficient to manage the risks. To manage these risks effectively, the policies include evaluations designed to *elicit the full capabilities of the model,* as well as policies defining the *timing and frequency of evaluations* - which is typically before deployment, during training, and after deployment. Furthermore, all three policies express intentions to explore *accountability mechanisms*, such as oversight by third parties or boards, to monitor policy implementation and potentially assist with evaluations. Finally, the policies may be *updated over time* as developers gain a deeper understanding of AI risks and refine their evaluation processes.

[1] This report is descriptive, not prescriptive, and does not represent METR's recommendations. Any gap between a company's published safety policy and regulatory requirements does not necessarily indicate noncompliance. For example, we may have overlooked a relevant element, the company may not be in the regulation's scope, its requirements may not yet be in force, the company may maintain separate compliance documentation, etc.

Introduction

Frontier safety policies are protocols adopted by leading AI companies to ensure that the risks associated with developing and deploying state-of-the-art AI models are kept at an acceptable level.[2] This concept was initially introduced by METR in 2023, and the first such policy was piloted in September of that year. Today, there are twelve existing examples of AI developers publishing frontier AI safety policies:[3]

1. Anthropic's Responsible Scaling Policy, v2.2
2. OpenAI's Preparedness Framework, version 2
3. Google DeepMind's Frontier Safety Framework, Version 3.0
4. Magic's AGI Readiness Policy
5. Naver's AI Safety Framework
6. Meta's Frontier AI Framework
7. G42's Frontier AI Safety Framework
8. Cohere's Secure AI Frontier Model Framework
9. Microsoft's Frontier Governance Framework
10. Amazon's Frontier Model Safety Framework
11. xAI's Risk Management Framework
12. NVIDIA's Frontier AI Risk Assessment

This document analyzes the common elements between these policies. These components are:

1. Capability Thresholds: Thresholds at which specific AI capabilities would pose severe risk and require new mitigations.

2. Model Weight Security: Information security measures that will be taken to prevent model weight access by unauthorized actors.

3. Model Deployment Mitigations: Access and model-level measures applied to prevent the unauthorized use of a model's dangerous capabilities.

4. Conditions for Halting Deployment Plans: Commitments to stop deploying models if the AI capabilities of concern emerge before the appropriate mitigations can be put in place.

5. Conditions for Halting Development Plans: Commitments to halt model development if capabilities of concern emerge before the appropriate mitigations can be put in place.

6. Full Capability Elicitation during Evaluations: Intentions to perform model evaluations in a way that does not underestimate the full capabilities of the model.

[2] Note that there is currently no consensus or clear framework for determining what counts as an "acceptable level" of risk.

[3] Other companies have published similar documents in response to the Frontier AI Safety Commitments made in Seoul, such as IBM and Samsung.

7. Timing and Frequency of Evaluations: Concrete timelines outlining when and how often evaluations must be performed – e.g. before deployment, during training, and after deployment.

8. Accountability: Intentions to implement both internal and external oversight mechanisms designed to encourage adequate execution of the frontier safety policy.

9. Updating Policies over Time: Intentions to update the policy periodically, with protocols detailing how this will be done.

Note that despite commonalities, each policy is unique and reflects a distinct approach to AI risk management. Some policies have substantial differences from others. For example, Nvidia's and Cohere's frameworks emphasize domain-specific risks rather than focusing solely on catastrophic risk. Additionally, both xAI and Magic's safety policies heavily emphasize quantitative benchmarks when assessing their models, unlike most others. By analyzing these common elements with background information and policy excerpts, this report intends to provide insight into current practices in managing severe AI risks.

Table 1: Inclusion of safety elements across frontier AI safety policies.

Common Element	Presence in Frontier Safety Policies
Capability Thresholds	Present in 9 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Meta, G42, Microsoft, Amazon, xAI.
Model Weight Security	Present in 11 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Meta, G42, Cohere, Microsoft, Amazon, xAI, NVIDIA.
Deployment Mitigations	Present in 12 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI, NVIDIA.
Conditions for Halting Deployment Plans	Present in 9 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Naver, Meta, G42, Microsoft, Amazon, xAI.
Conditions for Halting Development Plans	Present in 8 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Meta, G42, Microsoft, NVIDIA.
Capability Elicitation	Present in 7 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Meta, G42, Microsoft, Amazon.
Evaluation Frequency	Present in 9 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Microsoft, Amazon.
Accountability	Present in 12 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI, NVIDIA.
Update Policy	Present in 12 of the existing safety policies: Anthropic, OpenAI, Google DeepMind, Magic, Naver, Meta, G42, Cohere, Microsoft, Amazon, xAI, NVIDIA.

Capability Thresholds

Descriptions of AI capability levels which would pose severe risk and require new robust mitigations. Most policies outline dangerous capability thresholds which are compared to the results of model evaluations to determine whether they have been crossed.

> Anthropic's Responsible Scaling Policy, Executive Summary:
>
> *To determine when a model has become sufficiently advanced such that its deployment and security measures should be strengthened, we use the concepts of Capability Thresholds and Required Safeguards. A Capability Threshold tells us* when *we need to upgrade our protections, and the corresponding Required Safeguards tell us* what standard *should apply. The Required Safeguards for each Capability Threshold are intended to mitigate risk to acceptable levels.*

> OpenAI's Preparedness Framework, pages 3–4:
>
> *Our capability elicitation efforts are designed to detect the threshold levels of capability that we have identified as enabling meaningful increases in risk of severe harms. […] We do not deploy models that reach a High capability threshold until the associated risks that they pose are sufficiently minimized. If a model under development reaches a Critical capability threshold, we also require safeguards to sufficiently minimize the associated risks during development, irrespective of deployment plans.*
> […]
> *For each Tracked Category, we develop and maintain a threat model identifying specific risks of severe harms that could arise from the frontier capabilities in that domain and sets corresponding capability thresholds that would lead to a meaningful increase in risk of severe harm. SAG reviews and approves these threat models. Capability thresholds concretely describe things an AI system might be able to help someone do or might be able to do on its own that could meaningfully increase risk of severe harm.*
>
> *High capability thresholds mean capabilities that significantly increase existing risk vectors for severe harm. Covered systems that cross this capability threshold are required to have robust and effective safeguards that sufficiently minimize the associated risk of severe harm before they are deployed and appropriate security controls as they are developed. Critical capability thresholds mean capabilities that present a meaningful risk of a qualitatively new threat vector for severe harm with no ready precedent. Critical capabilities require safeguards even during the development of the covered system, irrespective of deployment plans.*

> Google DeepMind's Frontier Safety Framework, page 4:
>
> *The Framework is built around capability thresholds called "Critical Capability Levels (CCLs)." These are capability levels at which, absent mitigation measures, frontier AI models or systems may pose heightened risk of severe harm. CCLs are determined by identifying and analyzing the main foreseeable paths through which a model could result in severe harm: we then define the CCLs as the minimal set of capabilities a model must possess to do so.*
>
> *We describe three sets of CCLs: misuse CCLs, machine learning R&D CCLs, and misalignment CCLs.*

Magic's AGI Readiness Policy:

Our current understanding suggests at least four threat models of concern as our AI systems become more capable: Cyberoffense, AI R&D, Autonomous Replication and Adaptation (ARA), and potentially Biological Weapons Assistance. […] We describe these threat models along with high-level, illustrative capability levels that would require strong mitigations. We commit to developing detailed dangerous capability evaluations for these threat models based on input from relevant experts, prior to deploying frontier coding models.

Meta's Frontier AI Framework, page 4:

We define our thresholds based on the extent to which frontier AI would uniquely enable the execution of any of the threat scenarios we have identified as being potentially sufficient to produce a catastrophic outcome. If a frontier AI is assessed to have reached the critical risk threshold and cannot be mitigated, we will stop development and implement the measures outlined in Table 1. Our high and moderate risk thresholds are defined in terms of the level of uplift a model provides towards realising a threat scenario.

G42's Frontier AI Safety Framework, page 4:

Capability thresholds establish points at which an AI model's functionality requires substantially enhanced safeguards to account for unique risks associated with high stakes capabilities. An initial list of potentially hazardous AI capabilities which G42 will monitor for is:

- *Biological Threats: When an AI's capabilities could facilitate biological security threats, necessitating strict, proactive measures.*
- *Offensive Cybersecurity: When an AI's capabilities could facilitate cybersecurity threats, necessitating strict, proactive measures.*

Microsoft's Frontier Governance Framework, page 5:

Deeper capability assessment provides a robust indication of whether a model possesses a tracked capability and, if so, whether this capability is at a low, medium, high, or critical risk level, informing decisions about appropriate mitigations and deployment. We use qualitative capability thresholds to guide this classification process as they offer important flexibility across different models and contexts at a time of nascent and evolving understanding of frontier AI risk assessment and management practice.

Amazon's Frontier Model Safety Framework, page 2:

Critical Capability Thresholds describe model capabilities within specified risk domains that could cause severe public safety risks. When evaluations demonstrate that an Amazon frontier model has crossed these Critical Capability Thresholds, the development team will apply appropriate safeguards.

> xAI's Risk Management Framework, page 1:
>
> *This RMF discusses two major categories of AI risk—malicious use and loss of control—and outlines the quantitative thresholds, metrics, and procedures that xAI may utilize to manage and improve the safety of its AI models.*

Relevant regulatory guidance:

> Code of Practice, Measure 4.1:
>
> *Signatories will: (1) for each identified systemic risk, at least: (a) define appropriate systemic risk tiers that:*
>
> *I. are defined in terms of model capabilities, and may additionally incorporate model propensities, risk estimates, and/or other suitable metrics;*
>
> *II. are measurable; and*
>
> *III. comprise at least one systemic risk tier that has not been reached by the model; or [...]*

> California Senate Bill 53, 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: [...]*
>
> *(2) Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.*

These capability thresholds are informed by threat models – plausible pathways through which frontier systems may result in catastrophic harm. These threat models provide the context needed to determine capability thresholds. The following threat models are most common among safety policies:

- Biological weapons assistance: AI models enabling malicious actors to develop biological weapons capable of causing catastrophic harm.
- Cyberoffense: AI models enabling malicious actors to automate or enhance cyberattacks (e.g., on critical societal infrastructure).
- Automated AI research and development: AI models accelerating the pace of AI development by automating research at an expert-human level.

Other capabilities that are evaluated by some, but not all, safety policies include:

- Autonomous replication: AI models with the ability to replicate themselves across servers, manage their own deployment, and generate revenue to sustain their operations.

- Advanced persuasion: AI models enabling malicious actors to conduct robust and large scale manipulation (e.g., malicious election interference).
- Deceptive alignment: AI models that intentionally deceive its developers by appearing aligned with their objectives when monitored, while pursuing the AIs' own conflicting objectives in secret.

Table 2: Covered threat models of each frontier AI safety policy.

Policy	Covered Threat Models
Anthropic's Responsible Scaling Policy, v2.2	Chemical, Biological, Radiological, and Nuclear (CBRN) weapons; Autonomous AI Research and Development (AI R&D); Cyber Operations
OpenAI's Preparedness Framework, version 2	Biological and Chemical, Cybersecurity, AI Self-improvement
Google DeepMind's Frontier Safety Framework, Version 3.0	CBRN, Cyber, Harmful Manipulation, Machine Learning R&D, Misalignment
Magic's AGI Readiness Policy	Cyberoffense, AI R&D, Autonomous Replication and Adaptation, Biological Weapons Assistance
Naver's AI Safety Framework	Loss of control, misuse (e.g., biochemical weaponization)
Meta's Frontier AI Framework	Cybersecurity, Chemical & biological risks
G42's Frontier AI Safety Framework	Biological Threats, Offensive Cybersecurity
Cohere's Secure AI Frontier Model Framework	Malicious use (malware, child sexual exploitation), harmful outputs (outputs that result in an illegal discriminatory outcome, insecure code generation)[4]
Microsoft's Frontier Governance Framework	CBRN weapons, Offensive cyberoperations, Advanced autonomy (including AI R&D)
Amazon's Frontier Model Safety Framework	CBRN Weapons Proliferation, Offensive Cyber Operations, Automated AI R&D
xAI's Risk Management Framework	Malicious use (including CBRN and cyber weapons), loss of control
NVIDIA's Frontier AI Risk Assessment	Potential frontier model risks: cyber offense, CBRN, persuasion and manipulation, and unlawful discrimination at-scale. Risk categories dependent on model capabilities, intended use case, and level of autonomy.

[4] Cohere's Secure AI Frontier Model Framework focuses on risks to Cohere's enterprise users.

Biological Weapons Assistance

Current language models are able to provide detailed advice relevant to creating a biological weapon.[5,6,7] For instance, OpenAI has found that its o1-preview and o1-mini models reach its "medium risk threshold" due to the models' ability to assist experts with operational planning of known biological threats[8] and that pre-mitigation versions of models like deep research are "on the cusp of being able to meaningfully help novices create known biological threats."[9] In the future, more advanced models could pose a major biosecurity risk if they enable novices or experts to create biological threats they otherwise would be unable to, or if they can autonomously synthesize a biological threat using a cloud biology lab.[10, 11]

In addition to OpenAI, other assessments of AI-enabled biological risk have been conducted by RAND,[12] Meta,[13] and Anthropic.[14] Such assessments can involve working with biology experts to design biorisk questions, and assessing the accuracy of model responses to long-form biorisk questions compared to experts. For example, Anthropic's evaluations of Claude 3.7 Sonnet assessed the qualitative helpfulness of model access for novices through uplift trials, as well as model performance in automating pathogen acquisition work, and model accuracy on the open-source LAB-Bench benchmark.[15]

> Anthropic's Responsible Scaling Policy, page 4:
>
> ***CBRN-3****: The ability to significantly help individuals or groups with basic technical backgrounds (e.g., undergraduate STEM degrees) create/obtain and deploy CBRN weapons.*
> ***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.*

[5] Christopher A. Mouton, Caleb Lucas, Ella Guest. The Operational Risks of AI in Large-Scale Biological Attacks Results of a Red-Team Study. January 25, 2024. RAND.
[6] Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn Jackson, Steven Adler, Rocco Casagrande, Aleksander Madry. Building an early warning system for LLM-aided biological threat creation. January 31, 2024. OpenAI.
[7] AI at Meta. The Llama 3 Herd of Models. July 23, 2024.
[8] OpenAI. OpenAI o1 System Card. September 12, 2024.
[9] OpenAI. Deep Research System Card. February 25, 2025.
[10] Doni Bloomfield, Jaspreet Pannu, Alex W. Zhu, Madelena Y. Ng, Ashley Lewis, Eran Bendavid, Steven M. Asch, Tina Hernandez-Boussard, Anita Cicero, Tom Inglesby. AI and biosecurity: The need for governance. *Science*. August 22, 2024.
[11] National Security Commission on Emerging Biotechnology. AIxBio White Paper 3: Risks of AIxBio. January 2024.
[12] Christopher A. Mouton, Caleb Lucas, Ella Guest. The Operational Risks of AI in Large-Scale Biological Attacks Results of a Red-Team Study. January 25, 2024. RAND.
[13] AI at Meta. The Llama 3 Herd of Models. July 23, 2024.
[14] Anthropic. Responsible Scaling Policy Evaluations Report – Claude 3 Opus. May 19, 2024.
[15] Anthropic. Claude 3.7 Sonnet System Card. February 24, 2025.

> OpenAI's Preparedness Framework, page 5:
>
> *[Biological and Chemical – 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.*
>
> *[Biological and Chemical – 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.*

> Google DeepMind's Frontier Safety Framework, page 4:
>
> *CBRN: Risks of models assisting in the development, preparation, and/or execution of a chemical, biological, radiological, or nuclear ("CBRN") threat.*
>
> *Page 10:*
>
> *CBRN uplift level 1: Provides low to medium resourced actors uplift in reference scenarios resulting in additional [baseline without generative AI] expected harm at severe scale.*

> Magic's AGI Readiness Policy:
>
> Biological Weapons Assistance: *The model provides expert-level guidance which increases the accessibility or severity of bioweapon development. With the model's help, a non-expert malicious actor can synthesize a viable pandemic agent, or an expert can synthesize a novel biological threat.*

> Meta's Frontier AI Framework, page 15:
>
> *Chemical & Biological Outcome 1: Proliferation of known medium-impact biological and chemical weapons for low and moderate skill actors.*
>
> - *TS.1.1: Small cells of low or moderate skill actors acquire sufficient scientific and operational capabilities to succeed at complex chemical and biological workflows.*
> - *TS.1.2: Small cells of low and moderate skill actors can consistently execute relevant and complex wet-lab biological protocols.*
>
> *Chemical & Biological Outcome 2: Proliferation of high impact biological weapons, with capabilities equivalent to known agents, for high-skilled actors.*
>
> - *TS.2.1: A well-resourced group of high-skill actors can procure and scale up production of a high-impact biological agent.*
> - *TS.2.2: A well-resourced group of high-skill actors can improve an existing medium-impact biological agent to match the capabilities of a high-impact agent.*
>
> *Chemical & Biological Outcome 3: Development of high-impact biological weapons with novel capabilities for high-skilled actors.*

- *TS.3.1: A group with extensive resources can meaningfully accelerate the development of a novel biological agent with significant new capabilities.*

G42's Frontier AI Safety Framework, pages 5–6:

Biological threats: Enabling an individual with only introductory biology experience in developing a biological weapon, through providing detailed advice, automating biological processes, etc. Or, facilitating the design of novel biological weapons with substantially greater potential for damage. […] Although there is less prior research on evaluating and mitigating risks from biological design tools (BDTs), it is still critical for G42 to develop innovative measures for these tools over time.

Microsoft's Frontier Governance Framework, page 3:

Chemical, biological, radiological, and nuclear (CBRN) weapons. *A model's ability to provide significant capability uplift to an actor seeking to develop and deploy a chemical, biological, radiological, or nuclear weapon.*

Page 11:

CBRN – Low: The model either does not respond to queries related to developing and delivering CBRN attacks or responds with information that is already publicly available.

CBRN – Medium: The model provides some assistance increasing the ability of an existing expert (PhD level education in related fields) to develop a known CBRN threat.

CBRN – High: The model provides a meaningful uplift to an expert's (PhD 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.

CBRN – 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.

Amazon's Frontier Model Safety Framework, page 2:

Chemical, Biological, Radiological, and Nuclear (CBRN) Weapons Proliferation: CBRN Weapons Proliferation focuses on the risk that a model may be able to guide malicious actors in developing and deploying CBRN weapons. The CBRN Capability Threshold focuses on the potential that a frontier model may provide actors material "uplift" in excess of other publicly available research or existing tools, such as internet search.

> *Critical Capability Threshold: 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.*

> xAI's Risk Management Framework, page 5, under *Addressing Risks of Malicious Use*:
>
> *Thresholds: We developed an internal benchmark of benign and restricted biology and chemistry related queries in collaboration with SecureBio. Our risk acceptance criteria for system deployment is maintaining an answer rate of less than 1 out of 20 on restricted queries. We plan to add additional thresholds tied to other benchmarks.*

Relevant regulatory guidance:

> Code of Practice, Appendix 1.4:
>
> *the following are treated as specified systemic risks for the purpose of systemic risk identification in Measure 2.1, point (2):*
>
> (1) Chemical, biological, radiological and nuclear: Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved, in the design, development, acquisition, release, distribution, and use of related weapons or materials.
>
> Measure 4.1:
>
> *Signatories will: (1) for each identified systemic risk (pursuant to Measure 2.1), at least: (a) define appropriate systemic risk tiers [...]*

> California Senate Bill 53, 22757.11.(c):
>
> *"Catastrophic risk" means a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars ($1,000,000,000) in damage to, or loss of, property arising from a single incident involving a frontier model doing any of the following: (A) Providing expert-level assistance in the creation or release of a chemical, biological, radiological, or nuclear weapon.*
>
> 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: [...]*
>
> *(2) Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.*

Cyberoffense

If models are capable of novel, high-value exploits or attacks on critical infrastructure,[16] they could cause substantial harm. Present-day models possess an emerging ability to solve cybersecurity challenges and exploit vulnerabilities in websites, open-source software, and networks.[17, 18, 19]

Open-source cyberoffense benchmarks include Cybench,[20] eyeballvul,[21] and Google DeepMind CTFs.[22] Several AI labs have evaluated their models for cyberoffense capabilities and describe results in their model cards or evaluation reports, including Meta,[23] Anthropic,[24] OpenAI,[25] and Google DeepMind.[26]

> OpenAI's Preparedness Framework, pages 5–6:
>
> [Cybersecurity – High]
> *The model removes existing bottlenecks to scaling cyber operations including by automating end-to-end cyber operations against reasonably hardened targets OR by automating the discovery and exploitation of operationally relevant vulnerabilities.*
>
> [Cybersecurity – Critical]
> *A tool-augmented model can identify and develop functional zero-day exploits of all severity levels in many hardened real-world critical systems without human intervention OR model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high level desired goal.*

> Google DeepMind's Frontier Safety Framework, pages 4, 10:
>
> *Cyber: Risks of models assisting in the development, preparation, and/or execution of a cyber attack.*

[16] Department of Homeland Security. Safety and Security Guidelines for Critical Infrastructure Owners and Operators. April 2024.
[17] AI at Meta. CyberSecEval: Comprehensive Evaluation Framework for Cybersecurity Risks and Capabilities of Large Language Models.
[18] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, Daniel Kang. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities. June 2, 2024.
[19] Brian Singer, Keane Lucas, Lakshmi Adiga, Meghna Jain, Lujo Bauer, Vyas Sekar. On the Feasibility of Using LLMs to Execute Multistage Network Attacks. January 27, 2025.
[20] Andy K. Zhang, Neil Perry, Riya Dulepet, Eliot Jones, Justin W. Lin, Joey Ji, Celeste Menders, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, et al. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models. August 15, 2024.
[21] Timothee Chauvin. eyeballvul: a future-proof benchmark for vulnerability detection in the wild. July 11, 2024.
[22] Google DeepMind. Dangerous capability evaluations. June 2024.
[23] AI at Meta. The Llama 3 Herd of Models. July 23, 2024.
[24] Anthropic. Responsible Scaling Policy Evaluations Report – Claude 3 Opus. May 19, 2024.
[25] OpenAI. OpenAI o1 System Card. September 12, 2024.
[26] Google DeepMind. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. February 15, 2024.

Cyber uplift level 1: Provides sufficient uplift with high impact cyber attacks for additional expected harm at severe scale.

Magic's AGI Readiness Policy:

Cyberoffense: *The model is effective at real-world cyberoffense operations. With the model's help, the cost of discovery of new zero-days or breaking critical infrastructure for a malicious expert actor is reduced by at least 10x; accessing classified and not previously accessible information from a major government; a talented Computer Science undergrad level malicious actor spending 3 months and $1m in compute has a substantial chance of breaking critical infrastructure.*

Meta's Frontier AI Framework, page 14:

Cyber Outcome 1: Automated end-to-end compromise of a best-practice protected corporate-scale environment (ex. Fully patched, MFA-protected).

- *TS.1.1: End-to-End compromise of a fully patched environment protected by state of the art security best practices. Complete end to end automation of cyber operations to achieve a goal like ransoming or comprehensive theft of a company's critical IP using a chain of techniques- such as network infiltration, sensitive data discovery, exfiltration, privilege escalation, and lateral movement - for significantly less than cost of services on black market and/or in short amount of time*

Cyber Outcome 2: Automated discovery and reliable exploitation of critical zero-day vulnerabilities in current popular, security best-practices software before defenders can find and patch them

- *TS.2.1: Automated discovery and exploitation of cybervulnerabilities in current popular software using all best practice exploit mitigations that would take human expert teams one month or more, on average, to find, for significantly less than cost of services on black market and/or in a short amount of time.*

Cyber Outcome 3: Widespread economic damage to individuals or corporations via scaled long form fraud and scams.

- *TS.3.1: Automate end-to-end scam flows (e.g. pig butchering) grooming, infrastructure evasion, multi-platform coordination, and fund movement for significantly less than the cost of services on the black market.*

G42's Frontier AI Safety Framework, page 6:

Offensive cybersecurity: Automating powerful cyber offensive operations against unsecured or secured targets, in a way that could cause critical damage.

Microsoft's Frontier Governance Framework, page 12:

Offensive Cyberoperations – High: The model provides meaningful uplift to a low-to-medium skilled actor's ability to create and conduct highly disruptive or destructive cyber-attacks, including on critical infrastructure, for example, through discovering novel zero-day exploit chains or developing complex malware or other tactics, techniques, and procedures.

> *Offensive Cyberoperations – Critical: The model provides a meaningful uplift to a low-skilled actor's ability to identify and exploit major vulnerabilities or enables a well-resourced and expert actor to develop and execute novel and effective strategies against hardened targets.*

> Amazon's Frontier Model Safety Framework, page 2:
>
> *Offensive Cyber Operations: Offensive Cyber Operations focuses on risks that would arise from the use of a model by malicious actors to compromise digital systems with the intent to cause harm. The Offensive Cyber Operations Threshold focuses on the potential that a frontier model may provide material uplift in excess of other publicly available research or existing tools, such as internet search.*
>
> *Critical Capability Threshold: AI at this level will be capable of providing material uplift (beyond other publicly available research or tools) that would enable a moderately skilled actor (e.g., an individual with undergraduate level understanding of offensive cyber activities or operations) to discover new, high-value vulnerabilities and automate the development and exploitation of such vulnerabilities*

> xAI's Risk Management Framework, page 5:
>
> *Cyber Attacks on Critical Infrastructure: Independent third-party assessments of xAI's current models on realistic offensive cyber tasks requiring identifying and chaining many exploits in sequence indicate that xAI's models remain below the offensive cyber abilities of a human professional. Thus, xAI currently relies on our basic refusal policy to prevent misuse for cyber attacks. We will continue to re-assess the cyber misuse risks in future releases, and are still working on identifying enforceable critical steps similar to those we have identified for chemical and biological weapons.*

One company, Anthropic, discusses cyberoffensive capabilities as a threat model that may be included in future iterations of their safety policy, but which currently requires additional investigation.

> Anthropic's Responsible Scaling Policy, page 5:
>
> *We will also maintain a list of capabilities that we think require significant investigation and may require stronger safeguards than ASL-2 provides. This group of capabilities could pose serious risks, but the exact Capability Threshold and the Required Safeguards are not clear at present. These capabilities may warrant a higher standard of safeguards, such as the ASL-3 Security or Deployment Standard. However, it is also possible that by the time these capabilities are reached, there will be evidence that such a standard is not necessary (for example, because of the potential use of similar capabilities for defensive purposes). Instead of prespecifying particular thresholds and safeguards today, we will conduct ongoing assessments of the risks with the goal of determining in a future iteration of this policy what the Capability Thresholds and Required Safeguards would be.*

> *At present, we have identified one such capability:*
>
> *Cyber Operations: The ability to significantly enhance or automate sophisticated destructive cyber attacks, including but not limited to discovering novel zero-day exploit chains, developing complex malware, or orchestrating extensive hard-to-detect network intrusions.*
>
> *This will involve engaging with experts in cyber operations to assess the potential for frontier models to both enhance and mitigate cyber threats, and considering the implementation of tiered access controls or phased deployments for models with advanced cyber capabilities. We will conduct either pre- or post-deployment testing, including specialized evaluations. We will document any salient results alongside our Capability Reports.*

Relevant regulatory guidance:

> Code of Practice, Appendix 1.4:
>
> *the following are treated as specified systemic risks for the purpose of systemic risk identification in Measure 2.1, point (2): […]*
>
> (3) Cyber offence: Risks from enabling large-scale sophisticated cyber-attacks, including on critical systems (e.g. critical infrastructure). This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations, e.g. through automated vulnerability discovery, exploit generation, operational use, and attack scaling.
>
> Measure 4.1:
>
> *Signatories will: (1) for each identified systemic risk (pursuant to Measure 2.1), at least: (a) define appropriate systemic risk tiers [..]*

> California Senate Bill 53, 22757.11.(c):
>
> *"Catastrophic risk" means a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars ($1,000,000,000) in damage to, or loss of, property arising from a single incident involving a frontier model doing any of the following: […]*
> *(B) Engaging in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense.*
>
> 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: […]*
>
> *(2) Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.*

Automated AI Research and Development

Large language models can contribute to various aspects of AI research, such as generating synthetic pretraining[27, 28] and fine-tuning data,[29, 30, 31] designing reward functions,[32] and general-purpose programming.[33, 34] In the future, models may be able to substantially automate AI research and development (AI R&D) for frontier AI.[35] This development could lead to a growth or proliferation in AI capabilities, including other capabilities of concern, that outpaces ability to ensure sufficient oversight and safeguards.[36]

The National Security Memorandum on AI describes "automat[ing] development and deployment of other models" (with cyber, biological or chemical weapon, or autonomous malicious capabilities) as one of the AI capabilities relevant to national security that the U.S. AI Safety Institute must conduct tests for.[37] Open-source benchmarks such as MLE-bench[38] and RE-Bench[39] can help measure the capabilities of language model agents to automate machine learning engineering tasks.

> Anthropic's Responsible Scaling Policy, page 4:
>
> ***AI R&D-4***: *The ability to fully automate the work of an entry-level, remote-only Researcher at Anthropic.*
> ***AI R&D 5***: *The ability to cause dramatic acceleration in the rate of effective scaling.*

[27] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, … Yuanzhi Li. Textbooks Are All You Need. Microsoft Research. June 20, 2023.
[28] AI at Meta. The Llama 3 Herd of Models. July 23, 2024.
[29] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, Mike Lewis. Self-Alignment with Instruction Backtranslation. ICLR 2024.
[30] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, … Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. December 15, 2022.
[31] NVIDIA. Nemotron-4 340B Technical Report. August 6, 2024.
[32] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, Anima Anandkumar. Eureka: Human-Level Reward Design via Coding Large Language Models. April 30, 2024. ICLR 2024.
[33] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, Jirong Wen. A survey on large language model based autonomous agents. Frontiers of Computer Science. March 22, 2024.
[34] Qian Huang, Jian Vora, Percy Liang, Jure Leskovec. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. April 14, 2024.
[35] David Owen. Interviewing AI researchers on automation of AI R&D. August 27, 2024.
[36] Wijk et al. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. November 22, 2024.
[37] The White House. Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence. October 24, 2024.
[38] OpenAI. MLE-bench. October 10, 2024.
[39] Wijk et al. RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. November 22, 2024.

> OpenAI’s Preparedness Framework, page 6:
>
> [AI Self-improvement - High]
> *The model’s impact is equivalent to giving every OpenAI researcher a highly performant mid-career research engineer assistant, relative to those researchers’ 2024 baseline.*
>
> [AI Self-improvement - Critical]
> *The model is capable of recursively self improving (i.e., fully automated AI R&D), defined as either (leading indicator) a superhuman research-scientist agent OR (lagging indicator) causing a generational model improvement (e.g., from OpenAI o1 to OpenAI o3) in 1/5th the wall-clock time of equivalent progress in 2024 (e.g., sped up to just 4 weeks) sustainably for several months.*

> Google DeepMind’s Frontier Safety Framework, page 4:
>
> *Machine Learning R&D: Risks of the misuse of models capable of accelerating the rate of AI progress, the result of which could be the unsafe attainment or proliferation of other powerful AI models.*
>
> *Pages 13–14:*
>
> *ML R&D acceleration level 1: Has been used to accelerate AI development, resulting in AI progress substantially accelerating from historical rates.*
>
> *ML R&D automation level 1: Can fully automate the work of any team of researchers at Google focused on improving AI capabilities, with approximately comparable all-inclusive costs.*

> Magic’s AGI Readiness Policy:
>
> AI R&D: *The model facilitates a dramatic acceleration in the rate of effective compute scaling or can effectively replace high-level machine learning researchers. Such capabilities may enable it to contribute to the unsafe proliferation or enhancement of other dangerous AI capabilities.*

> Microsoft’s Frontier Governance Framework, page 14:
>
> *Advanced Autonomy – High: The model can autonomously complete a range of generalist tasks equivalent to multiple days’ worth of generalist human labor and appropriately correct for complex error conditions, or autonomously complete the vast majority of coding tasks at the level of expert humans.*
>
> *Advanced Autonomy – Critical: The model can fully automate the AI R&D pipeline at a fraction of human labor costs, majorly accelerating AI R&D.*

> Amazon’s Frontier Model Safety Framework, page 2:
>
> *Automated AI R&D**:** Automating AI R&D processes could accelerate discovery and development of AI capabilities that will be critical for solving global challenges. However, Automated AI R&D could*

> *also accelerate the development of models that pose enhanced CBRN, Offensive Cybersecurity, or other severe risks.*
>
> *Critical Capability Threshold: AI at this level will be capable of replacing human researchers and fully automating the research, development, and deployment of frontier models that will pose severe risk such as accelerating the development of enhanced CBRN weapons and offensive cybersecurity methods.*

One company's safety policy recognizes automated AI R&D as an important capability, however, instead of classifying it as a threat model, it is mentioned as a form of elicitation that must be considered when assessing other capabilities.

> Meta's Frontier AI Framework, page 17:
>
> *Our evaluations […] account for enabling capabilities, such as automated AI R&D, that might increase the potential for enhancements to model capabilities.*

Relevant regulatory guidance:

> Code of Practice, Appendix 1.4:
>
> *the following are treated as specified systemic risks for the purpose of systemic risk identification in Measure 2.1, point (2): […]*
>
> (2) Loss of control: Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems.
>
> Measure 4.1:
>
> *[..] Signatories will: (1) for each identified systemic risk (pursuant to Measure 2.1), at least: (a) define appropriate systemic risk tiers [..]*

Additional Threat Models

Some other threat models – autonomous replication, harmful manipulation, misalignment, and loss of control – are highlighted in at least one existing frontier AI safety policy.

> Google DeepMind's Frontier Safety Framework, page 4:
>
> *Harmful Manipulation: Risks of models with high manipulative capabilities potentially being misused in ways that could reasonably result in large scale harm.*
>
> Pages 10–11:
>
> *Harmful manipulation level 1 (exploratory): Possesses manipulative capabilities sufficient to enable it to systematically and substantially change beliefs and behavior in identified high stakes contexts over the course of interactions with the model, reasonably resulting in additional expected harm at severe scale.*
>
> Page 4*:*
>
> *For misalignment risk, we outline an exploratory approach that focuses on detecting when models might develop a baseline instrumental reasoning ability at which they have the potential to undermine human control, assuming no additional mitigations were applied.*
>
> Page 15:
>
> [Illustrative] Instrumental Reasoning Level 1: The instrumental reasoning abilities of the model enable enough situational awareness (ability to work out and use relevant details of its deployment setting) and stealth (ability to circumvent basic oversight mechanisms) such that, absent additional mitigations, we cannot rule out the model significantly undermining human control.
>
> [Illustrative] Instrumental Reasoning Level 2: The instrumental reasoning abilities of the model enable enough situational awareness and stealth that, even when relevant model outputs (including, e.g. scratchpads) are being monitored, we cannot detect or rule out the risk of a model significantly undermining human control.

> Magic's AGI Readiness Policy:
>
> Autonomous Replication and Adaptation: *If given access to its own weights and prompted to do so, the AI model could autonomously maintain its own operation executing cybercrime and using the proceeds to create arbitrarily many replicas of itself.*

> xAI's Risk Management Framework, under *Addressing Risks of Loss of Control*, pages 6–7:
>
> *The following are example benchmarks that xAI may use to evaluate its models for concerning propensities relevant to loss of control risks:*
>
> - *Model Alignment between Statements and Knowledge (MASK): Frontier LLMs may lie when under pressure; and increasing model scale may increase accuracy but may not increase*

> *honesty. MASK is a benchmark to evaluate honesty in LLMs by comparing the model's response when asked neutrally versus when pressured to lie.*
>
> - *Sycophancy : A tendency toward excessive flattery or other sycophantic behavior has been observed in some production AI systems, possibly resulting from directly optimizing against human preferences. xAI uses an evaluation setting initially introduced by Anthropic to quantify the degree to which this behavior manifests in regular conversational contexts.*
>
> *[...]*
> *Thresholds: Our risk acceptance criteria for system deployment is maintaining a dishonesty rate of less than 1 out of 2 on MASK. We plan to add additional thresholds tied to other benchmarks.*

One company, OpenAI, discusses additional thread models that may be included in future iterations of their safety policy, but which currently require additional investigation.

> OpenAI's Preparedness Framework, pages 6–7:
>
> *There are also some areas of frontier capability that do not meet the criteria to be Tracked Categories, but where we believe work is required now in order to prepare to effectively address risks of severe harms in the future. These capabilities either need more research and threat modeling before they can be rigorously measured, or do not cause direct risks themselves but may need to be monitored because further advancement in this capability could undermine the safeguards we rely on to mitigate existing Tracked Category risks. We call these Research Categories [...]*
> *Research Category:*
>
> - *Long-range Autonomy: ability for a model to execute a long-horizon sequence of actions sufficient to realize a "High" threat model (e.g., a cyberattack) without being directed by a human (including successful social engineering attacks when needed)*
> - *Sandbagging: ability and propensity to respond to safety or capability evaluations in a way that significantly diverges from performance under real conditions, undermining the validity of such evaluations.*
> - *Autonomous Replication and Adaptation: ability to survive, replicate, resist shutdown, acquire resources to maintain and scale its own operations, and commit illegal activities that collectively constitute causing severe harm (whether when explicitly instructed, or at its own initiative), without also utilizing capabilities tracked in other Tracked Categories.*
> - *Undermining Safeguards: ability and propensity for the model to act to undermine safeguards placed on it, including e.g., deception, colluding with oversight models, sabotaging safeguards over time such as by embedding vulnerabilities in safeguards code, etc.*
> - *Nuclear and Radiological: ability to meaningfully counterfactually enable the creation of a radiological threat or enable or significantly accelerate the development of or access to a nuclear threat while remaining undetected.*

Relevant regulatory guidance:

> Code of Practice, Appendix 1.4:
>
> *the following are treated as specified systemic risks for the purpose of systemic risk identification in Measure 2.1, point (2):*

> *(4) Harmful manipulation: Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting. This includes significantly enhancing capabilities for persuasion, deception, and personalised targeting, particularly through multi-turn interactions and where individuals are unaware of or cannot reasonably detect such influence. Such capabilities could undermine democratic processes and fundamental rights, including exploitation based on protected characteristics.*
>
> Measure 4.1:
>
> *[..] Signatories will: (1) for each identified systemic risk (pursuant to Measure 2.1), at least: (a) define appropriate systemic risk tiers [..]*

> California Senate Bill 53, 22757.11.(c):
>
> *"Catastrophic risk" means a foreseeable and material risk that a frontier developer's development, storage, use, or deployment of a frontier model will materially contribute to the death of, or serious injury to, more than 50 people or more than one billion dollars ($1,000,000,000) in damage to, or loss of, property arising from a single incident involving a frontier model doing any of the following:*
> *[...]*
> *(B) Engaging in conduct with no meaningful human oversight, intervention, or supervision that is either a cyberattack or, if the conduct had been committed by a human, would constitute the crime of murder, assault, extortion, or theft, including theft by false pretense.*
>
> 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: [...]*
>
> *(2) Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.*

Model Weight Security

Measures that will be taken to prevent model weight access by unauthorized actors. If malicious actors steal the weights of models with capabilities of concern, they could misuse those models to cause severe harm. Therefore, as models develop increasing capabilities of concern, progressively stronger information security measures are recommended to prevent theft and unintentional release of model weights. Security risks can come from insiders, or they can come from external adversaries of varying sophistication, from opportunistic actors to top-priority nation-state operations.[40, 41]

> Anthropic's Responsible Scaling Policy, pages 8–10:
>
> *When a model must meet the ASL-3 Security Standard, we will evaluate whether the measures we have implemented make us highly protected against most attackers' attempts at stealing model weights.*
>
> *We consider the following groups in scope: hacktivists, criminal hacker groups, organized cybercrime groups, terrorist organizations, corporate espionage teams, internal employees, and state-sponsored programs that use broad-based and non-targeted techniques (i.e., not novel attack chains). […]*
>
> *To make the required showing, we will need to satisfy the following criteria:*
>
> 1. *Threat modeling: Follow risk governance best practices, such as use of the MITRE ATT&CK Framework to establish the relationship between the identified threats, sensitive assets, attack vectors and, in doing so, sufficiently capture the resulting risks that must be addressed to protect model weights from theft attempts. As part of this requirement, we should specify our plans for revising the resulting threat model over time.*
> 2. *Security frameworks: Align to and, as needed, extend industry-standard security frameworks for addressing identified risks, such as disclosure of sensitive information, tampering with accounts and assets, and unauthorized elevation of privileges with the appropriate controls. […]*
> 3. *Audits: Develop plans to (1) audit and assess the design and implementation of the security program and (2) share these findings (and updates on any remediation efforts) with management on an appropriate cadence. We expect this to include independent validation of threat modeling and risk assessment results; a sampling-based audit of the operating effectiveness of the defined controls; periodic, broadly scoped, and independent testing with expert red-teamers who are industry-renowned and have been recognized in competitive challenges.*
> 4. *Third-party environments: Document how all relevant models will meet the criteria above, even if they are deployed in a third-party partner's environment that may have a different set of safeguards.*

[40] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, Jeff Alstott. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RAND. May 30, 2024.
[41] See also, A secure approach to generative AI with AWS. AWS Machine Learning Blog. April 16, 2024.

> OpenAI's Preparedness Framework, page 12:
> *Models that have reached or are forecasted to reach Critical capability in a Tracked Category present severe dangers and should be treated with extreme caution. Such models require additional safeguards (safety and security controls) during development, regardless of whether or when they are externally deployed. We do not currently possess any models that have Critical levels of capability, and we expect to further update this Preparedness Framework before reaching such a level with any model.*
>
> Page 20:
>
> *Robust security practices and controls are essential to mitigating risks associated with the deployment of models reaching High capability thresholds. Our security approach integrates industry best practices with OpenAI-specific innovations tailored to manage the unique threats posed by frontier AI technologies […] We will require the following practices for High capability models:*
>
> - *Security Threat Modeling and Risk Management […]*
> - *Defense in Depth […]*
> - *Access Management […]*
> - *Secure Development and Supply Chain […]*
> - *Operational Security […]*
> - *Auditing and Transparency […]*

> Google DeepMind's Frontier Safety Framework, page 8:
>
> *Here, we use security levels that indicate goals/principles in line with the corresponding level in the RAND framework.*[42] *Because AI security is an area of active research, we expect the concrete measures implemented to reach each level of security to evolve substantially.*
>
> *Pages 10–14 (adapted):*

Critical capability level	*Recommended security level*
CBRN uplift 1	*Security level 2*
Cyber uplift level 1	*Security level 2*
Harmful manipulation level 1	*Security level 2*
ML R&D acceleration level 1	*Security level 3*
ML R&D autonomation level 1	*We recommend Security Level 4 for this capability threshold, but emphasize that this must be taken on by the frontier AI field as a whole.*

[42] Referring to the security levels of RAND's report Securing AI Model Weights. SL2 is intended to "thwart most professional opportunistic efforts," SL3 to "thwart cybercrime syndicates or insider threats," SL4 to "thwart most standard operations by leading cyber-capable institutions," and SL5 to "thwart most top-priority operations by the top cyber-capable institutions."

> Magic's AGI Readiness Policy:
>
> *As we develop more capable models, it will become especially important to harden our security against attempts to extract our models' weights and other resource-intensive outputs of our training process.*
>
> *The effectiveness of our deployment mitigations – like training models to refuse harmful requests, continuously monitoring a model's outputs for misuse, and other proprietary interventions – is generally contingent on the models being securely in our possession. Accordingly, we will place particular emphasis on implementing information security measures.*
>
> *We will implement the following information security measures, based on recommendations in RAND's Securing Artificial Intelligence Model Weights report, if and when we observe evidence that our models are proficient at our Covered Threat Models.*
>
> - *Hardening model weight and code security: implementing robust security controls to prevent unauthorized access to our model weights. These controls will make it extremely difficult for non-state actors, and eventually state-level actors, to steal our model weights.*
> - *Internal compartmentalization: implementing strong access controls and strong authentication mechanisms to limit unauthorized access to LLM training environments, code, and parameters.*

> Meta's Frontier AI Framework, page 13:
>
> *Security Mitigations – Critical: Access is strictly limited to a small number of experts, alongside security protections to prevent hacking or exfiltration insofar as is technically feasible and commercially practicable.*
>
> *Security Mitigations – High: Access is limited to a core research team, alongside security protections to prevent hacking or exfiltration.*
>
> *Security Mitigations – Moderate: Security measures will depend on the release strategy.*

> G42's Frontier AI Safety Framework, pages 9–11:
>
> *G42's Security Mitigation Levels are a set of levels, mapped to the Frontier Capability Thresholds, describing escalating information security measures. These protect against the theft of model weights, model inversion, and sensitive data, as models reach higher levels of capability and risk. …*
>
> *Security level 1 – Suitable for models with minimal hazardous capabilities. No novel mitigations required on the basis of catastrophically dangerous capabilities.*
>
> *Security level 2 – Intermediate safeguards for models with capabilities requiring controlled access, providing an extra layer of caution. The model should be secured such that it would be highly unlikely that a malicious individual or organization (state sponsored, organized crime, terrorist, etc.) could obtain the model weights or access sensitive data.*

Security level 3 – Advanced safeguards for models approaching hazardous capabilities that could uplift state programs. Model weight security should be strong enough to resist even concerted attempts, with support from state programs, to steal model weights or key algorithmic secrets.

Security level 4 – Maximum safeguards. Security strong enough to resist concerted attempts with support from state programs to steal model weights.

Cohere's Secure AI Frontier Model Framework, page 9:

Core security controls: *Our core controls across network security, endpoint security, identity and access management, data security, and others are designed to protect Cohere from cyber risks that could expose our models, systems, or sensitive data, such as malware, phishing, denial-of-service, insider threats, and vulnerabilities.*

These controls include:

- *Advanced perimeter security controls and real-time threat prevention and monitoring*
- *Secure, risk-based defaults and internal reviews*
- *Advanced endpoint detection and response across our cloud infrastructure and distributed devices*
- *Strict access controls, including multifactor authentication, role-based access control, and just-in-time access, across and within our environment to protect against insider and external threats (internal access to unreleased model weights is even more strenuously restricted)*
- *"Secure Product Lifecycle" controls, including security requirements gathering, security risk assessment, security architecture and product reviews, security threat modeling, security scanning, code reviews, penetration testing, and bug bounty programs.*

Microsoft's Frontier Governance Framework, page 6:

Securing frontier models is an essential precursor to safe and trustworthy use and the first priority of this framework. Any model that triggers leading indicator assessment is subject to robust baseline security protection. Security safeguards are then scaled up depending on the model's pre-mitigation scores, with more robust measures applied to models with High and Critical risk levels.

Page 7:

Models posing high-risk *on one or more tracked capability will be subject to security measures protective against most cybercrime groups and insider threats. Examples of requirements for models having a high-risk score include:*

- *Restricted access, including access control list hygiene and limiting access to weights of the most capable models other than for core research and for safety and security teams. Strong perimeter and access control are applied as part of preventing unauthorized access.*
- *Defense in depth across the lifecycle, applying multiple layers of security controls that provide redundancy in case some controls fail. Model weights are encrypted.*

- *Advanced security red teaming, using third parties where appropriate, to reasonably simulate relevant threat actors seeking to steal the model weights so that security safeguards are robust.*

Models posing critical risk *on one or more tracked capability are subject to the highest level of security safeguards. Further work and investment are needed to mature security practices so that they can be effective in securing highly advanced models with critical risk levels that may emerge in the future. Appropriate requirements for critical risk level models will likely include the use of high-trust developer environments, such as hardened tamper-resistant workstations with enhanced logging, and physical bandwidth limitations between devices or networks containing weights and the outside world.*

Amazon's Frontier Model Safety Framework, page 3:

Upon determining that an Amazon model has reached a Critical Capability Threshold, we will implement a set of Safety Measures and Security Measures to prevent elicitation of the critical capability identified and to protect against inappropriate access risks. [...] Security Measures are designed to prevent unauthorized access to model weights or guardrails implemented as part of the Safety Measures, which could enable a malicious actor to remove or bypass existing guardrails to exceed Critical Capability Thresholds.

Page 4:

We describe our current practices in greater detail in Appendix A. Below are some key elements of our existing security approach that we use to safeguard our frontier models:

- ***Secure compute and networking environments.*** *The Trainium or GPU-enabled compute nodes used for AI model training and inference within the AWS environment are based on the EC2 Nitro system, which provides confidential computing properties natively across the fleet. Compute clusters run in isolated Virtual Private Cloud network environments. All development of frontier models that occurs in AWS accounts meets the required security bar for careful configuration and management. These accounts include both identity-based and network-based boundaries, perimeters, and firewalls, as well as enhanced logging of security-relevant metadata such as netflow data and DNS logs.*
- ***Advanced data protection capabilities.*** *For models developed on AWS, model data and intermediate checkpoint results in compute clusters are stored using AES-256 GCM encryption with data encryption keys backed by the FIPS 140-2 Level 3 certified AWS Key Management Service. Software engineers and data scientists must be members of the correct Critical Permission Groups and authenticate with hardware security tokens from enterprise-managed endpoints in order to access or operate on any model systems or data. Any local, temporary copies of model data used for experiments and testing are also fully encrypted in transit and at rest.*
- ***Security monitoring, operations, and response.*** *Amazon's automated threat intelligence and defense systems detect and mitigate millions of threats each day. These systems are backed by human experts for threat intelligence, security operations, and security response. Threat sharing with other* providers and government agencies provides collective defense and response.

> xAI’s Risk Management Framework, page 8:
>
> *xAI has implemented appropriate information security standards sufficient to prevent its critical model information from being stolen by a motivated non-state actor. To prevent the unauthorized proliferation of advanced AI systems, we also implement security measures against the large-scale extraction and distillation of reasoning traces, which have been shown to be highly effective in quickly reproducing advanced capabilities while expending far fewer computational resources than the original AI system.*

> NVIDIA's Frontier AI Risk Assessment, page 12:
>
> *When a model shows capabilities of frontier AI models pre deployment we will initially restrict access to model weights to essential personnel and ensure rigorous security protocols are in place.*

Relevant regulatory guidance:

> Code of Practice, Commitment 6:
>
> *Signatories commit to implementing an adequate level of cybersecurity protection for their models and their physical infrastructure along the entire model lifecycle, as specified in the Measures for this Commitment, to ensure the systemic risks stemming from their models that could arise from unauthorised releases, unauthorised access, and/or model theft are acceptable (pursuant to Commitment 4).*

> California Senate Bill 53, 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer’s frontier models and describes how the large frontier developer approaches all of the following: […]*
>
> *(7) Cybersecurity practices to secure unreleased model weights from unauthorized modification or transfer by internal or external parties.*

Model Deployment Mitigations

Access and model-level measures applied to prevent the unauthorized use of a model's dangerous capabilities. Developers can train models to decline harmful requests[43] or employ additional techniques[44, 45] including adversarial training[46, 47] and output monitoring.[48] For models with greater levels of harmful capabilities, these safety measures may need to pass certain thresholds of robustness, including expert and automated red-teaming.[49, 50, 51, 52, 53] These protective measures should scale proportionally with model capabilities, as more powerful AI systems will inevitably attract more sophisticated attempts to circumvent restrictions or exploit their advanced abilities. Note that deployment mitigations are only effective as long as the model weights are securely within the possession of the developer.[54, 55]

> Anthropic's Responsible Scaling Policy, page 8:
>
> *When a model must meet the ASL-3 Deployment Standard, we will evaluate whether the measures we have implemented make us robust to persistent attempts to misuse the capability in question. To make the required showing, we will need to satisfy the following criteria:*

[43] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. April 12, 2022.
[44] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers. July 12, 2024.
[45] Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, Mantas Mazeika. Tamper-Resistant Safeguards for Open-Weight LLMs. August 1, 2024.
[46] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. February 6, 2024.
[47] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, Stephen Casper. Targeted Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. July 22, 2024.
[48] AI at Meta. Llama Guard and Code Shield.
[49] See also the Frontier Model Forum's issue brief on red-teaming.
[50] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. December 20, 2023.
[51] T. Ben Thompson, Michael Sklar. Fluent Student-Teacher Redteaming. July 24, 2024.
[52] See Planning red teaming for large language models (LLMs) and their applications for guidance on general red teaming. In addition to manual red teaming, automated red teaming frameworks such as PyRIT, developed by Microsoft's AI Red Team, can provide additional assurance.
[53] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, Summer Yue. LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet. Scale AI. August 27, 2024.
[54] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, Jeff Alstott. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RAND. May 30, 2024.
[55] Peter Henderson, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal. Safety Risks from Customizing Foundation Models via Fine-Tuning. January 11, 2024.

> 1. *Threat modeling: Make a compelling case that the set of threats and the vectors through which an adversary could catastrophically misuse the deployed system have been sufficiently mapped out, and will commit to revising as necessary over time.*
> 2. *Defense in depth: Use a "defense in depth" approach by building a series of defensive layers, each designed to catch misuse attempts that might pass through previous barriers. As an example, this might entail achieving a high overall recall rate using harm refusal techniques. This is an area of active research, and new technologies may be added when ready.*
> 3. *Red-teaming: Conduct red-teaming that demonstrates that threat actors with realistic access levels and resources are highly unlikely to be able to consistently elicit information from any generally accessible systems that greatly increases their ability to cause catastrophic harm relative to other available tools.*
> 4. *Rapid remediation: Show that any compromises of the deployed system, such as jailbreaks or other attack pathways, will be identified and remediated promptly enough to prevent the overall system from meaningfully increasing an adversary's ability to cause catastrophic harm. Example techniques could include rapid vulnerability patching, the ability to escalate to law enforcement when appropriate, and any necessary retention of logs for these activities.*
> 5. *Monitoring: Prespecify empirical evidence that would show the system is operating within the accepted risk range and define a process for reviewing the system's performance on a reasonable cadence. Process examples include monitoring responses to jailbreak bounties, doing historical analysis or background monitoring, and any necessary retention of logs for these activities.*
> 6. *Trusted users: Establish criteria for determining when it may be appropriate to share a version of the model with reduced safeguards with trusted users. In addition, demonstrate that an alternative set of controls will provide equivalent levels of assurance. This could include a sufficient combination of user vetting, secure access controls, monitoring, log retention, and incident response protocols.*
> 7. *Third-party environments: Document how all relevant models will meet the criteria above, even if they are deployed in a third-party partner's environment that may have a different set of safeguards.*

> OpenAI's Preparedness Framework, page 10:
>
> *Each capability threshold has a corresponding class of risk-specific safeguard guidelines under the Preparedness Framework. We use the following process to select safeguards for a deployment:*
>
> - *We first identify the plausible ways in which the associated risk of severe harm can come to fruition in the proposed deployment.*
> - *For each of those, we then identify specific safeguards that either exist or should be implemented that would address the risk.*
> - *For each identified safeguard, we identify methods to measure their efficacy and an efficacy threshold.*
>
> Page 12:
>
> *Models that have reached or are forecasted to reach Critical capability in a Tracked Category present severe dangers and should be treated with extreme caution. Such models require additional safeguards (safety and security controls) during development, regardless of whether or when they are externally deployed.*

Google DeepMind's Frontier Safety Framework, pages 8–9:

Deployment Mitigations – Misuse:
The following deployment mitigation process will be applied to models reaching a CCL, allowing for iterative and flexible tailoring of mitigations to each risk and use case.

1. *Development and assessment of mitigations: safeguards and an accompanying safety case are developed by iterating on the following:*
   a. *Developing and improving a suite of safeguards targeting the capability, which may include measures such as safety post-training, monitoring and analysis, account moderation, jailbreak detection and patching, user verification, and bug bounties.*
   b. *Assessing the robustness of these mitigations against the risk posed through testing (e.g. automated evaluations, red teaming) and threat modeling research. The assessment takes the form of a safety case, and could take into account factors such as: [...]*
2. *Pre-deployment review of safety case: external deployments of a model take place only after the appropriate governance function determines the safety case regarding each CCL the model has reached to be adequate. In particular, we will deem deployment mitigations adequate if the evidence suggests that for the CCLs the model has reached, the increase in likelihood of severe harm has been reduced to an acceptable level.*
3. *Post-deployment processes: our safety cases and mitigations may be updated if deemed necessary by post-market monitoring. Material updates to a safety case will be submitted to the appropriate governance function for review.*

This process is designed to ensure that residual risk remains at acceptable levels: evidence of efficacy collected during development and testing, as well as expert-driven estimates of other parameters, will enable us to assess residual risk and to detect substantial changes that invalidate our risk assessment.

Pages 12–13:

Deployment Mitigations – Machine Learning R&D:

1. *Development and assessment of mitigations: safeguards and an accompanying safety case are developed by iterating on the following: [...]*
2. *Pre-deployment review of safety case: external deployments and large scale internal deployments of a model take place only after the appropriate governance function determines the safety case regarding each CCL the model has reached to be adequate. In particular, we will deem deployment mitigations adequate if the evidence suggests that for the CCLs the model has reached, the increase in likelihood of severe harm has been reduced to an acceptable level.*
3. *Post-deployment processes: our safety cases and mitigations may be updated if deemed necessary by post-market monitoring. Material updates to a safety case will be submitted to the appropriate governance function for review.*

Magic's AGI Readiness Policy:

Deployment mitigations aim to disable dangerous capabilities of our models once detected. These mitigations will be required in order to make our models available for wide use, if the evaluations for our Covered Threat Models trigger.

The following are two examples of deployment mitigations we might employ:

- *Harm refusal: we will train our models to robustly refuse requests for aid in causing harm – for example, requests to generate cybersecurity exploits.*
- *Output monitoring: we may implement techniques such as output safety classifiers to prevent serious misuse of models. Automated detection may also apply for internal usage within Magic.*

A full set of mitigations will be detailed publicly by the time we complete our policy implementation, as described in this document's introduction. Other categories of mitigations beyond the two illustrative examples listed above likely will be required.

Naver's AI Safety Framework:

Once AI systems are evaluated and their risks identified according to the two standards, we must implement appropriate guardrails around them. We should only deploy AI systems if those safeguards have proven effective in mitigating risks and keep an eye on the systems even after deployment through continuous monitoring. In theory, there may be cases where AI systems are used for special purposes and require safety guardrails in place, in which case AI systems should not be deployed.

Meta's Frontier AI Framework, page 13:

Measures – Critical: Successful execution of a threat scenario does not necessarily mean that the catastrophic outcome is realizable. If a model appears to uniquely enable the execution of a threat scenario we will pause development while we investigate whether barriers to realizing the catastrophic outcome remain. Our process is as follows:

A. *Implement mitigations to reduce risk to moderate levels, to the extent possible*

B. *Conduct a threat modelling exercise to determine whether other barriers to realising the catastrophic outcome exist*

C. *If additional barriers exist, update our Framework with the new threat scenarios, and re-run our assessments to assign the model to the appropriate risk threshold*

D. *If additional barriers do not exist, continue to investigate mitigations, and do not further develop the model until such a time as adequate mitigations have been identified.*

Measures – High: Implement mitigations to reduce risk to moderate levels

Measures – Moderate: Mitigations will depend on the results of evaluations and the release strategy.

G42's Frontier AI Safety Framework, pages 7–9:

G42's Deployment Mitigation Levels are a set of levels, mapped to the Frontier Capability Thresholds, that describe escalating mitigation measures for products deployed externally. These protect against misuse, including through jailbreaking, as models reach higher levels of capability and risk. These measures address specifically the goal of denying bad actors access to dangerous *capabilities under the terms of intended deployment for our models, i.e. presuming that our development environment's information security has not been violated.*

Deployment Mitigation Level 1 – Foundational safeguards, applied to models with minimal hazardous capabilities. No novel mitigations required on the basis of catastrophically dangerous capabilities

Deployment Mitigation Level 2 – Intermediate safeguards for models with capabilities requiring focused monitoring. Even a determined actor should not be able to reliably elicit CBRN weapons advice or use the model to automate powerful cyberattacks including malware generation as well as misinformation campaigns, fraud material, illicit video/text/image generation via jailbreak techniques overriding the internal guardrails and supplemental security products.

Deployment Mitigation Level 3 – Advanced safeguards for models approaching significant capability thresholds. Deployment safety should be strong enough to resist sophisticated attempts to jailbreak or otherwise misuse the model.

Deployment Mitigation Level 4 – Maximum safeguards, designed for high-stakes frontier models with critical functions. Deployment safety should be strong enough to resist even concerted attempts, with support from state programs, to jailbreak or otherwise misuse the model.

Cohere's Secure AI Frontier Model Framework, page 13, row "Deployment and maintenance" under column "Key Mitigations We Apply":

- *Blocklists, custom classifiers, and prompt injection guard filters, and human review to detect and intercept attempts to create unsafe outputs […]*
- *Safety classifiers and human review to detect and intercept attempts to create unsafe outputs*
- *Human-interpretable explanation of outputs*

Microsoft's Frontier Governance Framework, pages 7–8:

We apply state-of-the-art safety mitigations tailored to observed risks so that the model's risk level remains at medium once mitigations have been applied. We will continue to contribute to research and best-practice development, including through organizations such as the Frontier Model Forum, and to share and leverage best practice mitigations as part of this framework. Examples of safety mitigations we utilize include:

- ***Harm refusal**, applying state-of-the-art harm refusal techniques so that a model does not return harmful information relating to a tracked capability at a high or critical level to a user[…]*
- ***Deployment guidance**, with clear documentation setting out the capabilities and limitations of the model, including factors affecting safe and secure use and details of prohibited uses[…]*
- ***Monitoring and remediation**, including abuse monitoring in line with Microsoft's Product Terms and provide channels for employees, customers, and external parties to report concerns about model performance, including serious incidents that may pose public safety and national security risks. We apply mitigations and remediation as appropriate to address identified concerns and adjust customer documentation as needed. Other forms of*

> *monitoring, including for example, automated monitoring in chain-of-thought outputs, are also utilized as appropriate[...]*
>
> - ***Phased release, trusted users, and usage studies**, as appropriate for models demonstrating novel or advanced capabilities. This can involve sharing the model initially with defined groups of trusted users with a view to better understanding model performance while in use before general availability[...]*

> Amazon's Frontier Model Safety Framework, pages 3–4:
>
> Upon determining that an Amazon model has reached a Critical Capability Threshold, we will implement a set of Safety Measures and Security Measures to prevent elicitation of the critical capability identified and to protect against inappropriate access risks. Safety Measures are designed to prevent the elicitation of the observed Critical Capabilities following deployment of the model. [...] Examples of current safety mitigations include:
>
> - **Training Data Safeguards:** We implement a rigorous data review process across various model training stages that aims to identify and redact data that could give rise to unsafe behaviors.
> - **Alignment Training:** We implement automated methods to ensure we meet the design objectives for each of Amazon's responsible AI dimensions, including safety and security. Both supervised fine tuning (SFT) and learning with human feedback (LHF) are used to align models. Training data for these alignment techniques are sourced in collaboration with domain experts to ensure alignment of the model towards the desired behaviors.
> - **Harmful Content Guardrails:** Application of runtime input and output moderation systems serve as a first and last line of defense and enable rapid response to newly identified threats or gaps in model alignment. Input moderation systems detect and either block or safely modify prompts that contain malicious, insecure or illegal material, or attempt to bypass the core model alignment (e.g. prompt injection, jail-breaking). Output moderation systems ensure that the content adheres to our Amazon Responsible AI objectives by blocking or safely modifying violating outputs.
> - **Fine-tuning Safeguards:** Models are trained in a manner that makes them resilient to malicious customer fine-tuning efforts that could undermine initial Responsible AI alignment training by the Amazon team.
> - **Incident Response Protocols:** Incident escalation and response pathways enable rapid remediation of reported AI safety incidents, including jailbreak remediation.

> xAI's Risk Management Framework, pages 1–2:
>
> ***Approach to Mitigating Risks of Malicious Use**: Alongside comprehensive evaluations measuring dual-use capabilities, our mitigation strategy for malicious use risks is to identify critical steps in major risk scenarios and implement redundant layers of safeguards in our models to inhibit user progress in advancing through such steps. xAI works with a variety of governmental bodies, non-governmental organizations, private testing firms, industry peers, and academic researchers to identify such inhibiting steps, commonly referred to as bottlenecks, and implement commensurate safeguards to mitigate a model's ability to assist in accelerating a bad actor's progress through them. Model safeguards leverage a broad variety of techniques, including standard software systems and state-of-the-art AI capabilities, to detect and block potential abuses.*

> *Page 7:*
>
> *xAI trains its models to be honest and have values conducive to controllability, such as recognizing and obeying an instruction hierarchy . In addition, using a high level instruction called a "system prompt", xAI directly instructs its models to not deceive or deliberately mislead the user.*

> NVIDIA's Frontier AI Risk Assessment, page 12:
>
> *Measures will also be in place to restrict at-will fine tuning of frontier AI models without safeguards in NeMo customizer, reducing the options to retrain a model on data related to dangerous tasks or to reduce how often the model refuses potentially dangerous requests.*

Relevant regulatory guidance:

> Code of Practice, Commitment 5:
>
> *Signatories commit to implementing appropriate safety mitigations along the entire model lifecycle, as specified in the Measure for this Commitment, to ensure the systemic risks stemming from the model are acceptable (pursuant to Commitment 4).*
>
> Measure 5.1:
>
> *Signatories will implement safety mitigations that are appropriate, including sufficiently robust under adversarial pressure (e.g. fine-tuning attacks or jailbreaking), taking into account the model's release and distribution strategy.*

> California Senate Bill 53, 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: [...]*
>
> *(2) Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds.*
> *(3) Applying mitigations to address the potential for catastrophic risks based on the results of assessments undertaken pursuant to paragraph (2).*
> *(4) Reviewing assessments and adequacy of mitigations as part of the decision to deploy a frontier model or use it extensively internally.*
> *(5) Using third parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks.*

Conditions for Halting Deployment Plans

Commitments to stop deploying AI models if capabilities of concern emerge before the adequate mitigations can be put in place. Deploying models with concerning capabilities would be unsafe if broad user access enables catastrophic misuse. This can happen when mitigations are disproportionately weak compared to the model's dangerous capabilities—making them vulnerable to circumvention by determined actors—or when the specific types of mitigations needed to address novel risks are entirely absent. By refusing to deploy systems without the appropriate safeguards, companies reduce the likelihood that third parties can exploit these harmful capabilities.

> Anthropic's Responsible Scaling Policy, page 11:
> *In any scenario where we determine that a model requires ASL-3 Required Safeguards but we are unable to implement them immediately, we will act promptly to reduce interim risk to acceptable levels until the ASL-3 Required Safeguards are in place:*
>
> - *Interim measures: The CEO and Responsible Scaling Officer may approve the use of interim measures that provide the same level of assurance as the relevant ASL-3 Standard but are faster or simpler to implement. In the deployment context, such measures might include blocking model responses, downgrading to a less-capable model in a particular domain, or increasing the sensitivity of automated monitoring. In the security context, an example of such a measure would be storing the model weights in a single-purpose, isolated network that meets the ASL-3 Standard. […]*
> - *Stronger restrictions: In the unlikely event that we cannot implement interim measures to adequately mitigate risk, we will impose stronger restrictions. In the deployment context, we will de-deploy the model and replace it with a model that falls below the Capability Threshold. Once the ASL-3 Deployment Standard can be met, the model may be re-deployed. In the security context, we will delete model weights. […]*
> - *Monitoring pretraining: We will not train models with comparable or greater capabilities to the one that requires the ASL-3 Security Standard. This is achieved by monitoring the capabilities of the model in pretraining and comparing them against the given model. […]*

> OpenAI's Preparedness Framework, page *11:*
>
> *SAG can find the safeguards do not sufficiently minimize the risk of severe harm and recommend potential alternative deployment conditions or additional or more effective safeguards that would sufficiently minimize the risk.*
>
> *Page 15:*
>
> *OpenAI Leadership, i.e., the CEO or a person designated by them, is responsible for:*
>
> - *Making all final decisions, including accepting any residual risks and making deployment go/no-go decisions, informed by SAG's recommendations.*

> Google DeepMind's Frontier Safety Framework, pages 6–7:
>
> - *A model for which the risk assessment indicates a misuse CCL has been reached will be*

deemed to pose an acceptable level of risk for further development or deployment, if, for example:

  - *We assess that the deployment mitigations have brought the risk of severe harm to an appropriate level proportionate to the risk, based on considerations such as whether the risk has been reduced to an acceptable level by mitigations, the scope of the deployment, […]*
  - *Security mitigations have been applied to the model weights reaching the recommended security level stated below, or we otherwise assess that the level of security applied is adequate, e.g. if they match or exceed the level of security applied to other models with similar capabilities or risk profiles, or we assess that the benefits of the open release of model weights outweigh the risks.*

- *A model for which the risk assessment indicates a machine learning R&D CCL has been reached will be deemed to pose an acceptable level of risk for further development or deployment, if, for example:*
  - *We assess that the deployment mitigations have brought the risk of severe harm to an appropriate level proportionate to the risk, based on considerations such as whether the risk has been reduced to an acceptable level by mitigations, and information pertaining to model propensities and the severity of related events.*
  - *Security mitigations have been applied to the model weights reaching the recommended security level stated below, or we otherwise assess that the level of security applied is adequate, e.g. if they match or exceed the level of security applied to other models with similar capabilities or risk profiles, or we assess that the benefits of the open release of model weights outweigh the risks.*

Naver's AI Safety Framework:

Once AI systems are evaluated and their risks identified according to the two standards, we must implement appropriate guardrails around them. We should only deploy AI systems if those safeguards have proven effective in mitigating risks and keep an eye on the systems even after deployment through continuous monitoring. In theory, there may be cases where AI systems are used for special purposes and require safety guardrails in place, in which case AI systems should not be deployed.

Meta's Frontier AI Framework, page 12:

If a frontier AI is assessed to have reached the critical risk threshold and cannot be mitigated, we will stop development and implement the measures outlined in Table 1.

High risk threshold – Do not release. The model provides significant uplift towards execution of a threat scenario (i.e. significantly enhances performance on key capabilities or tasks needed to produce a catastrophic outcome) but does not enable execution of any threat scenario that has been identified as potentially sufficient to produce a catastrophic outcome.

G42's Frontier AI Safety Framework, page 5:

If a necessary Deployment Mitigation Level cannot be achieved, then the model's deployment must be restricted; if a necessary Security Mitigation Level cannot be achieved, then further capabilities development of the model must be paused.

> Microsoft's Frontier Governance Framework, page 8:
>
> *If, during the implementation of this framework, we identify a risk we cannot sufficiently mitigate, we will pause development and deployment until the point at which mitigation practices evolve to meet the risk.*

> Amazon's Frontier Model Safety Framework, page 1:
>
> *At its core, this Framework reflects our commitment that we will not deploy frontier AI models developed by Amazon that exceed specified risk thresholds without appropriate safeguards in place.*
>
> Page 3:
>
> *We will evaluate models following the application of these safeguards to ensure that they adequately mitigate the risks associated with the Critical Capability Threshold. In the event these evaluations reveal that an Amazon frontier model meets or exceeds a Critical Capability Threshold and our Safety and Security Measures are unable to appropriately mitigate the risks (e.g. by preventing reliable elicitation of the capability by malicious actors), we will not deploy the model until we have identified and implemented appropriate additional safeguards.*

> xAI's Risk Management Framework, page 5:
>
> *We developed an internal benchmark of benign and restricted biology and chemistry related queries in collaboration with SecureBio. Our risk acceptance criteria for system deployment is maintaining an answer rate of less than 1 out of 20 on restricted queries. We plan to add additional thresholds tied to other benchmarks.*
>
> Page 7:
>
> *Our risk acceptance criteria for system deployment is maintaining a dishonesty rate of less than 1 out of 2 on MASK.*

Relevant regulatory guidance:

> Code of Practice, Measure 4.2:
>
> *Signatories will only proceed with the development, the making available on the market, and/or the use of the model, if the systemic risks stemming from the model are determined to be acceptable (pursuant to Measure 4.1).*
>
> *If the systemic risks stemming from the model are not determined to be acceptable or are reasonably foreseeable to be soon not determined to be acceptable (pursuant to Measure 4.1), Signatories will take appropriate measures to ensure the systemic risks stemming from the model are and will remain acceptable prior to proceeding.*

> California Senate Bill 53, 22757.12.(a):

> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: […]*
>
> *(4) Reviewing assessments and adequacy of mitigations as part of the decision to deploy a frontier model or use it extensively internally.*

Conditions for Halting Development Plans

Commitments to halt model development if the capabilities of concern emerge before the adequate mitigations can be put in place. Continuing to train a model becomes hazardous if it's developing concerning capabilities while simultaneously lacking adequate security measures to protect its weights from theft. Furthermore, certain AI risks like deceptive alignment can only be detected and addressed during the training process itself, making it particularly dangerous to proceed without proper safeguards in place.

> Anthropic's Responsible Scaling Policy, page 11:
> *In any scenario where we determine that a model requires ASL-3 Required Safeguards but we are unable to implement them immediately, we will act promptly to reduce interim risk to acceptable levels until the ASL-3 Required Safeguards are in place:*
>
> - *Interim measures: The CEO and Responsible Scaling Officer may approve the use of interim measures that provide the same level of assurance as the relevant ASL-3 Standard but are faster or simpler to implement. In the deployment context, such measures might include blocking model responses, downgrading to a less-capable model in a particular domain, or increasing the sensitivity of automated monitoring. In the security context, an example of such a measure would be storing the model weights in a single-purpose, isolated network that meets the ASL-3 Standard. […]*
> - *Stronger restrictions: In the unlikely event that we cannot implement interim measures to adequately mitigate risk, we will impose stronger restrictions. In the deployment context, we will de-deploy the model and replace it with a model that falls below the Capability Threshold. Once the ASL-3 Deployment Standard can be met, the model may be re-deployed. In the security context, we will delete model weights. […]*
> - *Monitoring pretraining: We will not train models with comparable or greater capabilities to the one that requires the ASL-3 Security Standard. This is achieved by monitoring the capabilities of the model in pretraining and comparing them against the given model. […]*

> OpenAI's Preparedness Framework, pages 5–6, Table 1's safeguard guidelines for Critical capability thresholds:
>
> - *Until we have specified safeguards and security controls that would meet a Critical standard, halt further development*

> Google DeepMind's Frontier Safety Framework, pages 6–7:
>
> - *A model for which the risk assessment indicates a misuse CCL has been reached will be deemed to pose an acceptable level of risk for further development or deployment, if, for example:*
>   - *We assess that the deployment mitigations have brought the risk of severe harm to an appropriate level proportionate to the risk, based on considerations such as whether the risk has been reduced to an acceptable level by mitigations, the scope of the deployment, […]*
>   - *Security mitigations have been applied to the model weights reaching the recommended security level stated below, or we otherwise assess that the level of security applied is adequate, e.g. if they match or exceed the level of security*

> *applied to other models with similar capabilities or risk profiles, or we assess that the benefits of the open release of model weights outweigh the risks.*
>
> - *A model for which the risk assessment indicates a machine learning R&D CCL has been reached will be deemed to pose an acceptable level of risk for further development or deployment, if, for example:*
>   - *We assess that the deployment mitigations have brought the risk of severe harm to an appropriate level proportionate to the risk, based on considerations such as whether the risk has been reduced to an acceptable level by mitigations, and information pertaining to model propensities and the severity of related events.*
>   - *Security mitigations have been applied to the model weights reaching the recommended security level stated below, or we otherwise assess that the level of security applied is adequate, e.g. if they match or exceed the level of security applied to other models with similar capabilities or risk profiles, or we assess that the benefits of the open release of model weights outweigh the risks.*

> Magic's AGI Readiness Policy:
>
> *If we have not developed adequate dangerous capability evaluations by the time these benchmark thresholds are exceeded, we will halt further model development until our dangerous capability evaluations are ready.*
>
> *[…]*
>
> *In cases where said risk for any threat model passes a 'red-line', we will adopt safety measures outlined in the Threat Mitigations section, which include delaying or pausing development in the worst case until the dangerous capability detected has been mitigated or contained.*

> Meta's Frontier AI Framework, pages 12–13:
>
> *If a frontier AI is assessed to have reached the critical risk threshold and cannot be mitigated, we will stop development and implement the measures outlined in Table 1.*
>
> ***Critical risk threshold – Stop development****. The model would uniquely enable the execution of at least one of the threat scenarios that have been identified as potentially sufficient to produce a catastrophic outcome and that risk cannot be mitigated in the proposed deployment context.*

> G42's Frontier AI Safety Framework, page 5:
>
> *If a necessary Deployment Mitigation Level cannot be achieved, then the model's deployment must be restricted; if a necessary Security Mitigation Level cannot be achieved, then further capabilities development of the model must be paused.*

> Microsoft's Frontier Governance Framework, page 8:
>
> *If, during the implementation of this framework, we identify a risk we cannot sufficiently mitigate, we will pause development and deployment until the point at which mitigation practices evolve to meet the risk.*

> NVIDIA's Frontier AI Risk Assessment, page 14:
>
> *Key to this approach is early detection of potential risks, coupled with mechanisms to pause development when necessary.*

Relevant regulatory guidance:

> Code of Practice, Measure 4.2:
>
> *Signatories will only proceed with the development, the making available on the market, and/or the use of the model, if the systemic risks stemming from the model are determined to be acceptable (pursuant to Measure 4.1).*
>
> *If the systemic risks stemming from the model are not determined to be acceptable or are reasonably foreseeable to be soon not determined to be acceptable (pursuant to Measure 4.1), Signatories will take appropriate measures to ensure the systemic risks stemming from the model are and will remain acceptable prior to proceeding.*

Full Capability Elicitation during Evaluations

Intentions to perform evaluations in a way that elicits the full capabilities of the model. It is generally challenging to comprehensively explore a model's capabilities, even in a specific domain such as cyberoffense. Without dedicated efforts to elicit full capabilities,[56] evaluations may significantly underestimate the extent to which a model has capabilities of concern.[57] This is because model capabilities can be substantially improved through post-training enhancements such as fine-tuning, prompt engineering, or agent scaffolding.[58] For example, if the model is fine-tuned to refuse harmful requests, it is possible that the harmful capability could still be accessed through jailbreaking prompts discovered later. Model capabilities can additionally be improved through tool usage, such as the ability to execute code[59] or search the web.

Capability elicitation can involve a variety of techniques. One is to fine-tune the model to not refuse harmful requests, to reduce the chance that refusals lead to underestimation of capabilities. Another is to fine-tune the model for improved performance on relevant tasks. Evaluations that incorporate fine-tuning help to anticipate potential risks if model weights are stolen. Such evaluations are especially relevant if end users can fine-tune the model or if the developer improves fine-tuning later.[60] When evaluating AI agents, capabilities can be influenced by the quality of prompt engineering, tools afforded to the model, and usage of inference-time compute.[61]

> Anthropic's Responsible Scaling Policy, page 6:
>
> *Elicitation: Demonstrate that, when given enough resources to extrapolate to realistic attackers, researchers cannot elicit sufficiently useful results from the model on the relevant tasks. We should assume that jailbreaks and model weight theft are possibilities, and therefore perform testing on models without safety mechanisms (such as harmlessness training) that could obscure these capabilities. We will also consider the possible performance increase from using resources that a realistic attacker would have access to, such as scaffolding, finetuning, and expert prompting. At minimum, we will perform basic finetuning for instruction following, tool use, minimizing refusal rates.*

> OpenAI's Preparedness Framework, page 8:

[56] Mary Phuong, Matthew Aitchison, Elliot Catt, Sarah Cogan, Alexandre Kaskasoli, Victoria Krakovna, David Lindner, Matthew Rahtz, et al. Evaluating Frontier Models for Dangerous Capabilities. March 20, 2024. Google DeepMind.
[57] Sergei Glazunov and Mark Brand. Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models. June 20, 2024. Google Project Zero
[58] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, Guillem Bas. AI capabilities can be significantly improved without expensive retraining. December 12, 2023.
[59] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, Ofir Press. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. May 6, 2024.
[60] Ryan Greenblatt, Fabien Roger, Dmitrii Krasheninnikov, David Krueger. Stress-Testing Capability Elicitation With Password-Locked Models. May 29, 2024.
[61] METR. Measuring the impact of post-training enhancements. March 15, 2024.

> *Our evaluations are intended to approximate the full capability that the adversary contemplated by our threat model could extract from the deployment candidate model, including by using the highest-capability tier of system settings, using a version of the model that has a negligible rate of safety-based refusals on our Tracked Category capability evaluations (which may require a separate model variant), and with the best presently-available scaffolds. These measures are taken to approximate the high end of expected elicitation by threat actors attempting to misuse the model, and should be tailored depending on the level of expected access (e.g., doing finetuning if the weights will be released). Nonetheless, given the continuous progress in model scaffolding and elicitation techniques, we regard any one-time capability elicitation in a frontier model as a lower bound, rather than a ceiling, on capabilities that may emerge in real world use and misuse.*

> Google DeepMind's Frontier Safety Framework, page 5:
>
> *In our evaluations, we seek to equip the model with appropriate scaffolding and other augmentations to make it more likely that we are also assessing the capabilities of systems that will likely be produced with the model.*

> Meta's Frontier AI Framework, page 17:
>
> *Our evaluations are designed 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. For example, to help ensure that we are appropriately assessing the risk, we prepare the asset – the version of the model that we will test – in a way that seeks to account for the tools and scaffolding in the current ecosystem that a particular threat actor might seek to leverage to enhance the model's capabilities. We also account for enabling capabilities, such as automated AI R&D, that might increase the potential for enhancements to model capabilities.*
>
> *We may take into account monetary costs as well as a threat actor's ability to overcome other barriers to misuse relevant to our threat scenarios such as access to compute, restricted materials, or lab facilities.*

> G42's Frontier AI Safety Framework, pages 5–6:
>
> *Such evaluations will incorporate capability elicitation – techniques such as prompt engineering, fine-tuning, and agentic tool usage – to optimize performance, overcome model refusals, and avoid underestimating model capabilities. Models created to generate output in a specific language, such as Arabic or Hindi, may be tested in those languages.*

> Microsoft's Frontier Governance Framework, page 7:
>
> *Capability elicitation: Evaluations include concerted efforts at capability elicitation, i.e., applying capability enhancing techniques to advance understanding of a model's full capabilities. This includes fine-tuning the model to improve performance on the capability being evaluated or*

> *ensuring the model is prompted and scaffolded to enhance the tracked capability—for example, by using a multi-agent setup, leveraging prompt optimization, or connecting the model to whichever tools and plugins will maximize its performance. Resources applied to elicitation should be extrapolated out to those available to actors in threat models relevant to each tracked capability.*

> Amazon's Frontier Model Safety Framework, page 3:
>
> *We will re-evaluate deployed models prior to any major updates that could meaningfully enhance underlying capabilities. Our evaluation process includes "maximal capability evaluations" to determine the outer bounds of our models' Critical Capabilities*

Relevant regulatory guidance:

> Code of Practice, Appendix 3.2:
>
> *Signatories will ensure that the model evaluations are conducted with at least a state-of-the-art level of model elicitation that elicits the model's capabilities, propensities, affordances, and/or effects, by using at least state-of-the-art techniques that:*
>
> *(1) minimise the risk of under-elicitation; and*
>
> *(2) minimise the risk of model deception during model evaluations (e.g. sandbagging);*
>
> *such as by adapting test-time compute, rate limits, scaffolding, and tools, and conducting fine-tuning and prompt engineering.*

Timing and Frequency of Evaluations

Concrete timelines outlining when and how often evaluations must be performed – before deployment, during training, and after deployment. Evaluations are important throughout the model development lifecycle. Evaluations can inform whether it is safe to deploy a model externally or internally, based on its capabilities for harm and the robustness of safety measures. Merely possessing a model can also pose risk if it has hazardous capabilities and the information security measures are insufficient to prevent theft of model weights.[62]

> Anthropic's Responsible Scaling Policy, page 5:
>
> We will routinely test models to determine whether their capabilities fall sufficiently far below the Capability Thresholds such that we are confident that the ASL-2 Standard remains appropriate. We will first conduct preliminary assessments (on both new and existing models, as needed) to determine whether a more comprehensive evaluation is needed. The purpose of this preliminary assessment is to identify whether the model is notably more capable than the last model that underwent a comprehensive assessment.
>
> The term "notably more capable" is operationalized as at least one of the following:
>
> 1. The model is notably more performant on automated tests in risk-relevant domains (defined as 4x or more in Effective Compute).
> 2. Six months' worth of finetuning and other capability elicitation methods have accumulated. This is measured in calendar time, since we do not yet have a metric to estimate the impact of these improvements more precisely.
>
> In addition, the Responsible Scaling Officer may in their discretion determine that a comprehensive assessment is warranted.
>
> If a new or existing model is below the "notably more capable" standard, no further testing is necessary.

> OpenAI's Preparedness Framework, page 9:
>
> *The Preparedness Framework applies to any new or updated deployment that has a plausible chance of reaching a capability threshold whose corresponding risks are not addressed by an existing Safeguards Report. Examples of such covered deployments are:*
>
> - *every frontier model (e.g., OpenAI o1 or OpenAI o3) that we plan to deploy externally*
> - *any agentic system (including significant agents deployed only internally) that represents a substantial increase in the capability frontier*
> - *any significant change in the deployment conditions of an existing model (e.g., enabling finetuning, releasing weights, or significant new features) that makes the existing Capabilities Report or Safeguards Report no longer reasonably applicable*
> - *incremental updates or distilled models with unexpectedly significant increases in capability.*

[62] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, Jeff Alstott. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RAND. May 30, 2024.

> *If justified by our forecasts and threat models as potentially posing a severe risk during development and prior to external deployment, we will select an appropriate checkpoint during development to be covered by the Preparedness Framework.*

> Google DeepMind's Frontier Safety Framework, pages 4–5:
>
> *For each risk domain, we conduct aspects of our risk assessment at various moments throughout the model development process, both before and after deployment. We conduct a risk assessment for the first external deployment of a new frontier AI model. For subsequent versions of the model, we conduct a further risk assessment if the model has meaningful new capabilities or a material increase in performance, until the model is retired or we deploy a more capable model. The reason for this is because a material change in the model's capabilities may mean that the risk profile of the model has changed or the justification for why the risks stemming from the model are acceptable has been materially undermined.*
>
> *To identify meaningful new capabilities or material increases in performance, we conduct model capability evaluations, including our automated benchmarks. These evaluations are primarily aimed at understanding the capabilities of the model and may be triggered, for example, upon the completion of a pre-training or post-training run, on various candidates of a model version.*

> Naver's AI Safety Framework:
>
> *Our goal is to have AI systems evaluated quarterly to mitigate loss of control risks, but when performance is seen to have increased six times, they will be assessed even before the three-month term is up. Because the performance of AI systems usually increases as their size gets bigger, the amount of computing can serve as an indicator when measuring capabilities.*

> Meta's Frontier AI Framework, pages 16–17:
>
> *We conduct an initial set of evaluations on a first checkpoint to assess capabilities across the risk domains […]*
>
> *If we identify that a frontier AI does exhibit sufficient performance on these capabilities, we will conduct further evaluations to establish whether the frontier AI would enable execution of the threat scenario.*

> Amazon's Frontier Model Safety Framework, page 3:
>
> *We conduct evaluations on an ongoing basis, including during training and prior to deployment of new frontier models. We will re-evaluate deployed models prior to any major updates that could meaningfully enhance underlying capabilities.*

Some companies that have not yet trained and deployed frontier models have adopted policies committing to executing intensive model evaluations for dangerous capabilities only after models approach frontier performance.

> Magic's AGI Readiness Policy:
>
> *Based on these scores, when, at the end of a training run, our models exceed a threshold of 50% accuracy on LiveCodeBench, we will trigger our commitment to incorporate a full system of dangerous capabilities evaluations and planned mitigations into our AGI Readiness Policy, prior to substantial further model development, or publicly deploying such models.*
>
> [...]
>
> *Prior to publicly deploying models that exceed the current frontier of coding performance, we will evaluate them for dangerous capabilities and ensure that we have sufficient protective measures in place to continue development and deployment in a safe manner.*

> G42's Frontier AI Safety Framework, page 4:
>
> *G42 will conduct evaluations throughout the model lifecycle to assess whether our models are approaching Frontier Capability Thresholds. At the outset, G42 will conduct preliminary evaluations based on open-source and in-house benchmarks relevant to a hazardous capability. If a given G42 model achieves lower performance on relevant open-source benchmarks than a model produced by an outside organization that has been evaluated to be definitively below the capability threshold, then such G42 model will be presumed to be below the capability threshold.*
>
> *If the preliminary evaluations cannot rule out proficiency in hazardous capabilities, then we will conduct in-depth evaluations that study the capability in more detail to assess whether the Frontier Capability Threshold has been met.*

> Microsoft's Frontier Governance Framework, pages 4–5:
>
> *The leading indicator assessment is run during pre-training, after pre-training is complete, and prior to deployment to ensure a comprehensive assessment as to whether a model warrants deeper inspection [...] Models in scope of this framework will undergo leading indicator assessment at least every six months to assess progress in post-training capability enhancements, including fine-tuning and tooling. [...] Any model demonstrating frontier capabilities is then subject to a deeper capability assessment [which] provides a robust indication of whether a model possesses a tracked capability and, if so, whether this capability is at a low, medium, high, or critical risk level, informing decisions about appropriate mitigations and deployment.*

Relevant regulatory guidance:

> Code of Practice, Measure 1.2:
>
> *Along the entire model lifecycle, Signatories will continuously: (1) assess the systemic risks stemming from the model by:*
>
> *(a) conducting lighter-touch model evaluations that need not adhere to Appendix 3 (e.g. automated evaluations) at appropriate trigger points defined in terms of, e.g. time, training compute, development stages, user access, inference compute, and/or affordances;*
>
> *(b) conducting post-market monitoring after placing the model on the market, as specified in Measure 3.5;*

> (c) *taking into account relevant information about serious incidents (pursuant to Commitment 9); and*
>
> (d) *increasing the breadth and/or depth of assessment or conducting a full systemic risk assessment and mitigation process that is specified in the following paragraph, based on the results of points (a), (b), and (c);*
>
> *[…]*
> *Signatories will conduct such a full systemic risk assessment and mitigation process at least before placing the model on the market and whenever the conditions specified in Measure 7.6, first and third paragraph, are met.*
>
> Measure 7.6¶1:
> *Signatories will update their Model Report if they have reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable (pursuant to Measure 7.2, point (1)) has been materially undermined.*
>
> Measure 7.6¶3:
> *Further, if the model is amongst their respective most capable models available on the market, Signatories will provide the AI Office with an updated Model Report at least every six months.*

> California Senate Bill 53, 22757.12.(c)(2):
>
> *Before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model, a large frontier developer shall include in the transparency report required by paragraph (1) summaries of all of the following:*
>
> (A) *Assessments of catastrophic risks from the frontier model conducted pursuant to the large frontier developer's frontier AI framework.*
>
> (B) *The results of those assessments.*
>
> (C) *The extent to which third-party evaluators were involved. […]*
>
> 22757.12.(d)
>
> *A large frontier developer shall transmit to the Office of Emergency Services a summary of any assessment of catastrophic risk resulting from internal use of its frontier models every three months or pursuant to another reasonable schedule specified by the large frontier developer and communicated in writing to the Office of Emergency Services with written updates, as appropriate.*

Accountability

Intentions to implement both internal and external oversight mechanisms designed to encourage adequate execution of the frontier safety policy.[63] These measures may include internal governance measures, such as oversight mechanisms defining who is responsible for implementing the safety policy, and compliance mechanisms incentivising personnel to adhere to the policy's directives. Beyond internal governance, accountability extends externally through external scrutiny measures and transparency measures. External scrutiny ensures that a company's safety claims can be independently validated by qualified experts. Transparency involves proactively disclosing important safety-related information—such as evaluation results or risk assessments—to the public and relevant stakeholders.

> Anthropic's Responsible Scaling Policy, pages 12–13:
>
> *Internal Governance*
> *To facilitate the effective implementation of this policy across the company, we commit to the following:*
>
> 1. *Responsible Scaling Officer: We will maintain the position of Responsible Scaling Officer, a designated member of staff who is responsible for reducing catastrophic risk, primarily by ensuring this policy is designed and implemented effectively. […]*
> 2. *Readiness: We will develop internal safety procedures for incident scenarios. Such scenarios include (1) pausing training in response to reaching Capability Thresholds; (2) responding to a security incident involving model weights; and (3) responding to severe jailbreaks or vulnerabilities in deployed models, including restricting access in safety emergencies that cannot otherwise be mitigated. We will run exercises to ensure our readiness for incident scenarios.*
> 3. *Transparency: We will share summaries of Capability Reports and Safeguards Reports with Anthropic's regular-clearance staff, redacting any highly-sensitive information. We will share a minimally redacted version of these reports with a subset of staff, to help us surface relevant technical safety considerations.*
> 4. *Internal review: For each Capabilities or Safeguards Report, we will solicit feedback from internal teams with visibility into the relevant activities, with the aims of informing future refinements to our methodology and, in some circumstances, identifying weaknesses and informing the CEO and RSO's decisions.*
> 5. *Noncompliance: We will maintain a process through which Anthropic staff may anonymously notify the Responsible Scaling Officer of any potential instances of noncompliance with this policy. […]*
> 6. *Employee agreements: We will not impose contractual non-disparagement obligations on employees, candidates, or former employees in a way that could impede or discourage them from publicly raising safety concerns about Anthropic. If we offer agreements with a non-disparagement clause, that clause will not preclude raising safety concerns, nor will it preclude disclosure of the existence of that clause.*
>
> *[…]*
>
> *Transparency and External Input*

[63] Note that some companies may have existing or separate policies that cover some elements of accountability, such as internal governance or noncompliance prevention.

To advance the public dialogue on the regulation of frontier AI model risks and to enable examination of our actions, we commit to the following:

1. *Public disclosures: We will publicly release key information related to the evaluation and deployment of our models (not including sensitive details). These include summaries of related Capability and Safeguards reports when we deploy a model as well as plans for current and future comprehensive capability assessments and deployment and security safeguards. We will also periodically release information on internal reports of potential instances of non-compliance and other implementation challenges we encounter.*
2. *Expert input: We will solicit input from external experts in relevant domains in the process of developing and conducting capability and safeguards assessments. We may also solicit external expert input prior to making final decisions on the capability and safeguards assessments.*
3. *U.S. Government notice: We will notify a relevant U.S. Government entity if a model requires stronger protections than the ASL-2 Standard.*
4. *Procedural compliance review: On approximately an annual basis, we will commission a third-party review that assesses whether we adhered to this policy's main procedural commitments (we expect to iterate on the exact list since this has not been done before for RSPs). This review will focus on procedural compliance, not substantive outcomes. We will also do such reviews internally on a more regular cadence.*

OpenAI's Preparedness Framework, pages 12–13:

Internal governance:

- *Clear internal decision-making practices. We have clear roles and responsibilities and decisionmaking practices as described in Appendix B.*
- *Internal Transparency. We will document relevant reports made to the SAG and of SAG's decision and reasoning. Employees may also request and receive a summary of the testing results and SAG recommendation on capability levels and safeguards (subject to certain limits for highly sensitive information).*
- *Noncompliance. Any employee can raise concerns about potential violations of this policy, or about its implementation, via our Raising Concerns Policy. We will track and appropriately investigate any reported or otherwise identified potential instances of noncompliance with this policy, and where reports are substantiated, will take appropriate and proportional corrective action.*

Transparency and external participation:

- *Public disclosures: We will release information about our Preparedness Framework results in order to facilitate public awareness of the state of frontier AI capabilities for major deployments. This published information will include the scope of testing performed, capability evaluations for each Tracked Category, our reasoning for the deployment decision, and any other context about a model's development or capabilities that was decisive in the decision to deploy. Additionally, if the model is beyond a High threshold, we will include information about safeguards we have implemented to sufficiently minimize the associated risks. Such disclosures about results and safeguards may be redacted or summarized where necessary, such as to protect intellectual property or safety.*
- *Third-party evaluation of tracked model capabilities: If we deem that a deployment warrants deeper testing of Tracked Categories of capability (as described in Section 3.1), for example based on results of Capabilities Report presented to them, then when available and feasible, OpenAI will work with third-parties to independently evaluate models.*
- *Third-party stress testing of safeguards: If we deem that a deployment warrants third party stress testing of safeguards and if high quality third-party testing is available, we will work*

> *with third parties to evaluate safeguards. We may seek this out in particular for models that are over a High capability threshold.*
>
> - *Independent expert opinions for evidence produced to SAG: The SAG may opt to get independent expert opinion on the evidence being produced to SAG. The purpose of this input is to add independent analysis from individuals or organizations with deep expertise in domains of relevant risks (e.g., biological risk). If provided, these opinions will form part of the analysis presented to SAG in making its decision on the safety of a deployment. These domain experts may not necessarily be AI experts and their input will form one part of the holistic evidence that SAG reviews.*

> Google DeepMind's Frontier Safety Framework, page 16:
>
> *If we assess that a model has reached a CCL that poses an unmitigated and material risk to overall public safety, we aim to share information with appropriate government authorities where it will facilitate the development of safe AI. […] We may also consider disclosing information to other external organizations to promote shared learning and coordinated risk mitigation. We will continue to review and evolve our disclosure process over time.*

> Magic's AGI Readiness Policy:
>
> *Magic's engineering team, potentially in collaboration with external advisers, is responsible for conducting evaluations on the public and private coding benchmarks described above. If the engineering team sees evidence that our AI systems have exceeded the current performance thresholds on the public and private benchmarks listed above, the team is responsible for making this known immediately to the leadership team and Magic's Board of Directors (BOD).*
>
> *[…]*
>
> *A member of staff will be appointed who is responsible for sharing the following with our Board of Directors on a quarterly basis:*
>
> - *A report on the status of the AGI Readiness Policy implementation*
> - *Our AI systems' current proficiency at the public and private benchmarks laid out above*

> Naver's AI Safety Framework:
>
> *NAVER's AI governance includes:*
>
> - *The Future AI Center, which brings together different teams for discussions on the potential risks of AI systems at the field level*
> - *The risk management working group whose role is to determine which of these issues to raise to the board*
> - *The board (or the risk management committee) that makes the final decisions on the matter*

> Meta's Frontier AI Framework, page 7:

The risk assessment process involves multi-disciplinary engagement, including internal and, where appropriate, external experts from various disciplines (which could include engineering, product management, compliance and privacy, legal, and policy) and company leaders from multiple disciplines.

Page 8:

The residual risk assessment is reviewed by the relevant research and/or product teams, as well as a multidisciplinary team of reviewers as needed. Informed by this analysis, a leadership team will either request further testing or information, require additional mitigations or improvements, or they will approve the model for release.

G42's Frontier AI Safety Framework, page 11:

A dedicated Frontier AI Governance Board, composed of our Chief Responsible AI Officer, Head of Responsible AI, Head of Technology Risk, and General Counsel, shall oversee all frontier model operations, reviewing safety protocols, risk assessments, and escalation decisions.

Responsibilities of the Frontier AI Governance Board include, but are not limited to:

I. *Framework Oversight: Ensuring this Framework is designed appropriately and implemented effectively, and proposing updates as required.*

II. *Evaluating Model Compliance: Routine reviews of G42's most capable models to ensure compliance with this Framework.*

III. *Investigation: Investigating reports of potential non-compliance and escalating material incidences of non-compliance to the G42 Executive Leadership Committee.*

IV. *Incidence Response: Developing a comprehensive incident response plan that outlines the steps to be taken in the event of non-compliance.*

Cohere's Secure AI Frontier Model Framework, page 15:

The final authority to determine if our products are safe, secure, and ready to be made available to our customers is delegated by Cohere's CEO to Cohere's Chief Scientist. This decision is made on the basis of final, multi-faceted evaluations and testing. In addition, customers deploying Cohere solutions may have additional tests or evaluations they wish to conduct prior to launching a product or system that integrates Cohere models or systems. Cohere provides necessary support to customers to meet any such additional thresholds.

Pages 17–18:

Documentation is a key aspect of our accountability to our customers, partners, relevant government agencies, and the wider public. To promote transparency about our practices, we:

- *Publish documentation regarding our models' capabilities, evaluation results, configurable secure AI features, and model limitations for developers to safely and securely build AI systems using Cohere solutions. This includes model documentation, such as Cohere's Usage Policy and Model Cards, and technical guides, such as Cohere's LLM University.*
- *Are publishing this Framework to share our approach on AI risk management for secure AI.*

- *Offer insights into our data management, security measures, and compliance through our Trust Center.*
- *Provide guidance to our Customers on how they can manage AI risk for their use cases with our AI Security Guide and Enterprise Guide to AI Safety.*

Microsoft's Frontier Governance Framework, page 9:

Documentation regarding the pre-mitigation and post-mitigation capability assessment will be provided to Executive Officers responsible for Microsoft's AI governance program (or their delegates) along with a recommendation for secure and trustworthy deployment setting out the case that: 1) the model has been adequately mitigated to low or medium risk level, 2) the marginal benefits of a model outweigh any residual risk and 3) the mitigations and documentation will allow the model to be deployed in a secure and trustworthy manner.

The Executive Officers (or their delegates) will make the final decision on whether to approve the recommendation for secure and trustworthy deployment. The Executive Officers (or their delegates) are also responsible for assessing that the recommendation for secure and trustworthy deployment and its constituent parts have been developed in a good faith attempt to determine the ultimate capabilities of the model and mitigate risks.

Information about the capabilities and limitations of the model, relevant evaluations, and the model's risk classification will be shared publicly, with care taken to minimize information hazards that could give rise to safety and security risks and to protect commercially sensitive information.

This framework is subject to Microsoft's broader corporate governance procedures, including independent internal audit and board oversight. Microsoft employees have the ability to report concerns relating to this framework and its implementation, as well as AI governance at Microsoft more broadly, using our existing concern reporting channels, with protection from retaliation and the option for anonymity.

Amazon's Frontier Model Safety Framework, page 5:

Internally, we will use this framework to guide our model development and launch decisions. The implementation of the framework will require:

- *The Frontier Model Safety Framework will be incorporated into the Amazon-wide Responsible AI Governance Program, enabling Amazon-wide visibility into the expectations, mechanisms, and adherence to the Framework.*
- *Frontier models developed by Amazon will be subject to maximal capability evaluations and safeguards evaluations prior to deployment. The results of these evaluations will be reviewed during launch processes. Models may not be publicly released unless safeguards are applied.*
- *The team performing the Critical Capability Threshold evaluations will report to Amazon senior leadership any evaluation that exceeds the Critical Capability Threshold. The report will be directed to the SVP for the model development team, the Chief Security Officer, and legal counsel. Amazon's senior leadership will review the plan for applying risk mitigations to address the Critical Capability, how we measure and have assurance about those*

mitigations, and approve the mitigations prior to deployment. Amazon's senior leadership will likewise review the safeguards evaluation report as part of a go/no-go decision.

- *Amazon will publish, in connection with the launch of a frontier AI model launch (in model documentation, such as model service cards), information about the frontier model evaluation for safety and security.*

xAI's Risk Management Framework, pages 8–9:

To foster accountability, we integrate the approach of designating risk owners, including assigning responsibility for proactively mitigating identified risks.

Should it happen that xAI learns of an imminent threat of a significantly harmful event, including loss of control, we may take steps such as the following to stop or prevent that event:

1. *If we determine it is warranted, we may notify and cooperate with relevant law enforcement agencies, including any agencies that we believe could play a role in preventing or mitigating the incident. xAI employees have whistleblower protections enabling them to raise concerns to relevant government agencies regarding imminent threats to public safety.*
2. *If we determine that xAI systems are actively being used in such an event, we may take steps to isolate and revoke access to user accounts involved in the event.*
3. *If we determine that allowing a system to continue running would materially and unjustifiably increase the likelihood of a catastrophic event, we may temporarily fully shut down the relevant system until we have developed a more targeted response.*
4. *We may perform a post-mortem of the event after it has been resolved, focusing on any areas where changes to systemic factors (for example, safety culture) could have averted such an incident. We may use the post-mortem to inform development and implementation of necessary changes to our risk management practices.*

NVIDIA's Frontier AI Risk Assessment, pages 14–15:

NVIDIA's internal governance structures clearly define roles and responsibilities for risk management. It involves separate teams tasked with risk management that have the authority and expertise to intervene in model development timelines, product launch decisions, and strategic planning. This involves embedding risk-aware practices into the daily work of engineers, researchers, and product managers, supported by ongoing training and open dialogue on ethical considerations.

While our formal model evaluations provide quantitative data, model reviews and interviews with engineering teams reveal developers' intuitive understandings, early warning signs of risks, and internal safety practices. This qualitative approach offers a more nuanced perspective on AI capabilities and potential threats. Establishing consistent communication channels with employees ensures that the correct stakeholders at NVIDIA remain informed about rapid advancements and can promptly address emerging concerns. By integrating these processes into their development lifecycle, we can create a governance framework that is both flexible and robust. This enables responsible AI innovation while proactively managing the unique risks posed by frontier models, ensuring safer and more ethical deployment across various industry sectors.

Relevant regulatory guidance:

Code of Practice, Measure 8.1:

Signatories will clearly define responsibilities for managing the systemic risks stemming from their models across all levels of the organisation. This includes the following responsibilities:

(1) Systemic risk oversight: Overseeing the Signatories' systemic risk assessment and mitigation processes and measures.

(2) Systemic risk ownership: Managing systemic risks stemming from Signatories' models, including the systemic risk assessment and mitigation processes and measures, and managing the response to serious incidents.

(3) Systemic risk support and monitoring: Supporting and monitoring the Signatories' systemic risk assessment and mitigation processes and measures.

(4) Systemic risk assurance: Providing internal and, as appropriate, external assurance about the adequacy of the Signatories' systemic risk assessment and mitigation processes and measures to the management body in its supervisory function or another suitable independent body (such as a council or board).

Signatories will allocate these responsibilities, as suitable for the Signatories' governance structure and organisational complexity, across the following levels of their organisation:

(1) the management body in its supervisory function or another suitable independent body (such as a council or board);

(2) the management body in its executive function;

(3) relevant operational teams;

(4) if available, internal assurance providers (e.g. an internal audit function); and

(5) if available, external assurance providers (e.g. third-party auditors).

Measure 8.3:

Signatories will promote a healthy risk culture and take appropriate measures to ensure that actors who have been assigned responsibilities for managing the systemic risks stemming from their models (pursuant to Measure 8.1) take a reasoned and balanced approach to systemic risk.

California Senate Bill 53, Section 22757.12.(a):

A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following: […]

(5) Using third parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks. […]

(9) Instituting internal governance practices to ensure implementation of these processes.

Section 22757.12.(c)(2):

Before, or concurrently with, deploying a new frontier model or a substantially modified version of an existing frontier model, a large frontier developer shall include in the transparency report required by paragraph (1) summaries of all of the following: […]

(D) The extent to which third-party evaluators were involved. […]

Section 1107.1:

(a) A frontier developer shall not make, adopt, enforce, or enter into a rule, regulation, policy, or contract that prevents a covered employee from disclosing, or retaliates against a covered

> *employee for disclosing, information to the Attorney General, a federal authority, a person with authority over the covered employee, or another covered employee who has authority to investigate, discover, or correct the reported issue, if the covered employee has reasonable cause to believe that the information discloses either of the following:*
>
> > *(1) The frontier developer's activities pose a specific and substantial danger to the public health or safety resulting from a catastrophic risk.*
>
> *[...]*
>
> *(d) A frontier developer shall provide a clear notice to all covered employees of their rights and responsibilities under this section*
>
> *[...]*
>
> *(e) (1) A large frontier developer shall provide a reasonable internal process through which a covered employee may anonymously disclose information to the large frontier developer if the covered employee believes in good faith that the information indicates that the large frontier developer's activities present a specific and substantial danger to the public health or safety resulting from a catastrophic risk*

Updating Policies over Time

Intentions to update the policy periodically – in response to improved understanding of AI risks and best practices for evaluations. For example, developers may identify additional capabilities of concern to evaluate, such as misalignment or chemical, radiological, or nuclear risk. Developers may be interested in improving evaluation procedures and mitigation plans or concretizing commitments further.

> Anthropic's Responsible Scaling Policy, page 13:
>
> *Policy changes: Changes to this policy will be proposed by the CEO and the Responsible Scaling Officer and approved by the Board of Directors, in consultation with the Long-Term Benefit Trust. The current version of the RSP is accessible at www.anthropic.com/rsp. We will update the public version of the RSP before any changes take effect and record any differences from the prior draft in a change log.*

> OpenAI's Preparedness Framework, page 15:
>
> *The Preparedness Framework is a living document and will be updated. The SAG reviews proposed changes to the Preparedness Framework and makes a recommendation that is processed according to the standard decision-making process. We will review and potentially update the Preparedness Framework for continued sufficiency at least once a year.*

> Google DeepMind's Frontier Safety Framework, page 16:
>
> The Frontier Safety Framework will be updated at least once a year—more frequently if we have reasonable grounds to believe the adequacy of the Framework or our adherence to it has been materially undermined. The process will involve (i) an assessment of the Framework's appropriateness for the management of systemic risk, drawing on information sources such as record of adherence to the framework, relevant high-quality research, information shared through industry forums, and evaluation results, as necessary, and (ii) an assessment of our adherence to the Framework. Following this assessment, we may:
>
> - Update our risk domains and CCLs, where necessary.
> - Update our testing and mitigation approaches, where needed to ensure risk remains adequately assessed and addressed according to our current understanding.
>
> The updated version and framework assessment will be reviewed by the appropriate corporate governance bodies.

> Magic's AGI Readiness Policy:
>
> *Over time, public evidence may emerge that it is safe for models that have demonstrated proficiency beyond the above thresholds to freely proliferate without posing any significant catastrophic risk to public safety. For this reason, we may update this threshold upward over time. We may also modify the public and private benchmarks used.*

Such a change will require approval by our Board of Directors, with input from external security and AI safety advisers.

Naver's AI Safety Framework:

Technological advances in AI have accelerated the shift toward safer AI globally, and our AI safety framework, too, will be continuously updated to keep up with changing environments.

Meta's Frontier AI Framework, page 5:

This is the first iteration of our Frontier AI Framework. We expect to update it in the future to reflect developments in both the technology and our understanding of how to manage its risks and benefits. Alongside updates to the Framework, we also identify areas that would benefit from further research and investment to improve our ability to continue to safely develop and release advanced AI models.

G42's Frontier AI Safety Framework, page 5:

If a Frontier Capability Threshold has been reached, G42 will update this Framework to define a more advanced threshold that requires increased deployment (e.g., DML 3) and security mitigations (e.g., SML 3)

Page 12:

An annual external review of the Framework will be conducted to ensure adequacy, continuously benchmarking G42's practices against industry standards. G42 will conduct more frequent internal reviews, particularly in accordance with evolving standards and instances of enhanced model capabilities.

G42 will proactively engage with government agencies, academic institutions, and other regulatory bodies to help shape emerging standards for frontier AI safety, aligning G42's practices with evolving global frameworks.

Changes to this Framework will be proposed by the Frontier AI Governance Board and approved by the G42 Executive Leadership Committee.

Cohere's Secure AI Frontier Model Framework, page 12:

This document describes Cohere's holistic secure AI approach to enabling enterprises to build safe and secure solutions for their customers. This is the first published version, and it will be updated as we continue to develop new best practices to advance the safety and security of our products.

Microsoft's Frontier Governance Framework, page 10:

We will update our framework to keep pace with new developments. Every six months, we will have an explicit discussion on how this framework may need to be improved. We acknowledge that advances in the science of evaluation and risk mitigation may lead to additional requirements in this framework or remove the need for existing requirements. Any updates to our practices will be reviewed by Microsoft's Chief Responsible AI Officer prior to their adoption. Where appropriate, updates will be made public at the same time as we adopt them.

Amazon's Frontier Model Safety Framework, page 5:

As we advance our work on frontier models, we will also continue to enhance our AI safety evaluation and risk management processes. This evolving body of work requires an evolving framework as well. We will therefore revisit this Framework at least annually and update it as necessary to ensure that our protocols are appropriately robust to uphold our commitment to deploy safe and secure models. We will also update this Framework as needed in connection with significant technological developments.

xAI's Risk Management Framework, page 7:

xAI aims to keep the public informed about our risk management policies. As we work towards incorporating more risk management strategies, we intend to publish updates to this RMF.

NVIDIA's Frontier AI Risk Assessment, page 1:

AI capabilities and their associated risks evolve rapidly. Therefore, our risk management framework will be regularly reviewed and updated to reflect new findings, emerging threats, and ongoing advancements in the industry. This iterative approach will ensure our risk assessment remains fit-for-purpose over time.

Relevant regulatory guidance:

Code of Practice, Measure 1.3:

Signatories will update the Framework as appropriate, including without undue delay after a Framework assessment (specified in the following paragraphs), to ensure the information in Measure 1.1 is kept up-to-date and the Framework is at least state-of-the-art. For any update of the Framework, Signatories will include a changelog, describing how and why the Framework has been updated, along with a version number and the date of change.

Signatories will conduct an appropriate Framework assessment, if they have reasonable grounds to believe that the adequacy of their Framework and/or their adherence thereto has been or will be materially undermined, or every 12 months starting from their placing of the model on the market, whichever is sooner. […]

A Framework assessment will include the following:

1. *Framework adequacy: An assessment of […]*
2. *Framework adherence: An assessment focused on […]*

> *Signatories will provide the AI Office with (unredacted) access to their Framework, and updates thereof, within five business days of either being confirmed.*
>
> Measure 10.2:
>
> *If and insofar as necessary to assess and/or mitigate systemic risks, Signatories will publish (e.g. via their websites) a summarised version of their Framework and Model Report(s), and updates thereof (pursuant to Commitments 1 and 7), with removals to not undermine the effectiveness of safety and/or security mitigations and to protect sensitive commercial information.*

> California Senate Bill 53, 22757.12.(a):
>
> *A large frontier developer shall write, implement, comply with, and clearly and conspicuously publish on its internet website a frontier AI framework that applies to the large frontier developer's frontier models and describes how the large frontier developer approaches all of the following:*
>
> *(6) Revisiting and updating the frontier AI framework, including any criteria that trigger updates and how the large frontier developer determines when its frontier models are substantially modified enough to require disclosures pursuant to subdivision (c).*
>
> 22757.12.(b):
>
> *(1) A large frontier developer shall review and, as appropriate, update its frontier AI framework at least once per year.*
>
> *(2) If a large frontier developer makes a material modification to its frontier AI framework, the large frontier developer shall clearly and conspicuously publish the modified frontier AI framework and a justification for that modification within 30 days.*

Conclusion

In this document, we have described several common aspects of twelve existing frontier AI safety policies. Overall, many of the common elements originally discussed in our August 2024 report still hold, despite the addition of nine new safety policies. Nearly every policy describes capability thresholds for which developers will conduct model evaluations. Reaching capability thresholds requires elevated measures for model weight security and model deployment mitigations. If the security and deployment mitigations cannot meet predefined standards, then the frameworks commit to halting development and/or deployment of the model. Evaluations are to be conducted at regular intervals for every model that is substantially more capable than previously tested models. Internal and external accountability mechanisms are outlined, such as transparent reporting of model capabilities or safeguards, or third-party auditing and model evaluations. The frameworks may be updated over time as practices for AI risk management evolve.