RESEARCH PAPER
JULY 2025

Mapping Industry Practices to the EU AI Act's GPAI Code of Practice Safety and Security Measures

Authors: Lily Stelling, Mick Yang, Rokas Gipiškis, Leon Staufer, Ze Shen Chin, Siméon Campos, Ariel Gil, Michael Chen

Mapping Industry Practices to the EU AI Act's GPAI Code of Practice Safety and Security Measures

Lily Stelling,[1*] Mick Yang,[2*] Rokas Gipiškis,[3,4*]
Leon Staufer,[5] Ze Shen Chin,[3,6] Siméon Campos,[7] Ariel Gil,[3] Michael Chen[8]

[1] University of Queensland, [2] University of Pennsylvania, [3] AI Standards Lab,
[4] Vilnius University, [5] Technical University of Munich, [6] Oxford Martin AI Governance Initiative, [7] SaferAI, [8] METR
* Equal contribution

Corresponding author: lilystelling1@gmail.com

Abstract

This report provides a detailed comparison between the Safety and Security measures proposed in the EU AI Act's General-Purpose AI (GPAI) Code of Practice (Third Draft) and the current commitments and practices voluntarily adopted by leading AI companies. As the EU moves toward enforcing binding obligations for GPAI model providers, the Code of Practice will be key for bridging legal requirements with concrete technical commitments. Our analysis focuses on the draft's Safety and Security section (Commitments II.1–II.16), documenting excerpts from current public-facing documents that are relevant to each individual measure.

We systematically reviewed different document types, such as companies' frontier safety frameworks and model cards, from over a dozen companies, including OpenAI, Anthropic, Google DeepMind, Microsoft, Meta, Amazon, and others. This report is not meant to be an indication of legal compliance, nor does it take any prescriptive viewpoint about the Code of Practice or companies' policies. Instead, it aims to inform the ongoing dialogue between regulators and General-Purpose AI model providers by surfacing evidence of industry precedent for various measures. Nonetheless, we were able to find relevant quotes from at least 5 companies' documents for the majority of the measures in Commitments II.1–II.16.

Executive Summary

This report was prepared during the ongoing Code of Practice finalization process and should be viewed as a snapshot of industry precedent and publicly available documentation as of mid-2025. The primary aim is to offer a timely evidence base to inform discussions among policymakers, industry stakeholders, and civil society. This report does *not* assess whether any specific company or model is *compliant* with the EU AI Act or the Code of Practice.

Out of 72 Measures,[1] our analysis indicates that 52 Commitments and Measures had at least 3 separate company documents that were found to have relevant quotes for each sub-section.[2] The majority of the measures we analysed reached saturation of 5 relevant quotes found. This suggests that the Codes of Practice mostly aligns with existing industry precedent, rather than imposing novel regulatory burdens.

We chose to list *up to 5 quotes* for each measure (and if applicable, sub-sections of measures). The table below lists the measures and relevant sub-sections which had *2 or less* quotes found. While a measure containing 3–5 relevant quotes by no means indicates full compliance of the measure—and a measure with 2 or fewer quotes may be due to our method missing suitable quotes—we provide this table, as it may still be informative to readers which measures had less coverage by company documents.[3]

[1] This includes Commitments and its Measures. This also excludes Commitment II.14 and its Measures.
[2] A sub-section is a term we use to refer to the numbered clauses or paragraphs within a given Measure. (This number was deduced by excluding the 20 measures in Table 1, i.e. 52 measures and/or commitments had at least 3 relevant quotes for each of its sub-section (if applicable).
[3] Note that Commitment II.14 had relatively few quotes, though we do not include it in the table as it refers specifically to notifying the EU AI Office. Since the Code of Practice has not been finalized at the time of writing, we do not feel the exclusion of quotes is informative here.

Measures, commitments and sub-sections with fewer than 3 relevant quotes found	
II.1.2, sub-section (3)	II.8.2, sub-section (2)
II.1.3, sub-sections (2) and (3)	II.8.3, sub-sections (3), (7), (9)–(11), (13)
II.1.5, sub-section (2)	II.8.4, sub-sections (2) and (3)
II.3.2	II.8.6, sub-sections (1)-(3)
II.4.9, sub-section (2)	II.9.3, sub-section (3)
II.4.10, sub-sections (1) and (2)	II.11.1, multiple parts
II.4.11, sub-sections (3)–(4)	II.12.2, multiple parts
II.4.12, sub-section (1)	II.12.3
II.4.14, sub-sections (1)–(10)	II.12.4
II.5.1	II.13

Table 1. Measures, commitments and sub-sections with fewer than 3 relevant quotes found.

Table of Contents

Introduction

Purpose and intended audience

This report compares current industry practices with the commitments outlined in the third draft of the General-Purpose AI (GPAI) Code of Practice, specifically the Safety and Security section. By examining each Safety and Security measure alongside relevant statements and policies from leading AI companies, we aim to provide evidence of industry precedent for each item to ground discussion of the Code of Practice's commitments.

This report is primarily intended for:

- Policymakers requiring an initial evidence base of industry precedent relating to the Code of Practice.
- Companies wanting to compare and build their corporate AI governance plans alongside peer practice.
- Researchers analyzing how governance topics in the Code of Practice are discussed in public company documents.

At the time of publication, the final Code of Practice has been published. However, since its Safety and Security section was streamlined from the Third Draft's, we believe our analysis is still relevant to the above intended audience for the same reasons.

Context: Code of Practice and industry precedent

Providers of GPAI models with systemic risk will be subject to Articles 52–55 of the AI Act.[4] The Code of Practice stipulated in Article 56 is expected to be the default means of demonstrating compliance once approved by the Commission. The Code of Practice's drafting process, which began after the AI Act's entry into force on August 1, 2024, aims to produce a finalized Code before the AI Act's GPAI provisions become effective in August 2025.

[4] *GPAI models*—also known as foundation models—that pose a *systemic risk* are the most advanced models at that point in time or of an equivalent impact. This is determined either on a case-by-case basis or by exceeding 10^25 FLOPs when training the model. (see more at Article 3(63) and 3(65)). A 'provider of GPAI models with systemic risk' refers to a natural or legal person, public authority, agency or other body, who develops such a model or has such a model developed and placed on the EU market.

Frontier AI companies have increasingly adopted voluntary safety policies and transparency mechanisms ahead of regulatory requirements. Most prominent among these are:[5]

- **Frontier AI safety policies**: Public company policies governing testing and risk assessment methodologies for training and deployment decisions.
- **Model cards**: Detailed documentation describing model capabilities, limitations, training data characteristics, and intended use cases.
- **Evaluation and red-teaming reports**: Reports on model performance against safety benchmarks and vulnerability testing.

Our approach

For each measure in the EU AI Act's draft GPAI Code of Practice (Commitments II.1–II.16), we selected quotes from industry documents that provide context for understanding how companies' current practice could apply to the Code of Practice's commitments.

- Companies covered were signatories of the AI Seoul Summit 2024 with publicly released frontier safety frameworks.[6]
- Sources included model/system cards, frontier AI safety frameworks, policy statements, and technical reports from major AI developers. We prioritized formal organization-wide safety frameworks first, then model cards or product release reports, then other statements.
- Quotes were chosen and sorted by relevance to specific measures (and if applicable, sub-sections of measures), with quotes from up to 5 different companies per sub-section maximum. For measures with more than five relevant industry quotes, we included only the top five most relevant quotes. We chose over-inclusion at first pass, then iterated on inclusion thresholds.
- Quality assurance: For robustness, each section was then assigned to a separate individual to review the chosen quotes and ensure the evidence selected was relevant for the corresponding item in the Code of Practice.[7]

[5] There are also a number of resources that analyze industry precedent of risk management, such as SaferAI's Frontier AI Risk Management Framework, the Frontier Model Forum's Components of Frontier AI Safety Frameworks, and METR's Common Elements of Frontier AI Safety Policy. However, these resources do not provide a direct comparison of existing company documents with the General-Purpose AI Code of Practice.

[6] We did not include: 01.AI, Inflection AI, Minimax, Mistral AI, Technology Innovation Institute, Zhipu AI, Samsung. As of writing, these companies had not yet published a safety framework. We did include IBM's Responsible Use Guide, but note that this may not be considered a safety framework as the other documents are, as it does not engage with severe risks, thresholds, and security controls for unreleased model weights.

[7] See Author contribution for details.

Results of our Analysis

Below, we consider each of the commitments and measures in the Safety and Security section (Commitments II.1–16) of the third draft of the General-Purpose AI Code of Practice for the EU AI Act, which is reproduced in full throughout this report.

Under each subheading, we quote the commitment or measure, along with relevant excerpts from public documents published by AI companies, shown in blue boxes. Documents include system cards, model cards, frontier AI safety commitments, AI research papers, and frontier AI safety policies. We focus on companies who have published frontier AI safety policies as of June 2025.

For robustness, after quotes were initially selected, we reviewed each others sections to ensure the evidence selected was indeed relevant for the corresponding item in the Code.

The quotes listed are not exhaustive evidence of industry precedent, and some quotes are more relevant than others. This means there could be some relevant quotes not listed here. We try to order quotes from most to least relevant for each item, but do not claim to have found the most relevant existing quotes. Rather, we intend this analysis to form a sample of existing evidence of industry precedent. We leave interpretation up to the reader on how closely aligned quotes are with the Code's proposed items.

Commitment II.1. Safety and Security Framework

> "Signatories commit to adopting and implementing a Safety and Security Framework (hereafter, "the Framework") that will: (1) apply to the Signatories' GPAISRs [general-purpose AIs with systemic risk]; and (2) detail the systemic risk assessment, systemic risk mitigation, and governance risk mitigation measures and procedures that Signatories intend to adopt to keep systemic risks stemming from their GPAISRs within acceptable levels."

> A number of AI companies have adopted frontier AI safety policies that commit to systemic risk assessment, systemic risk mitigation, and governance measures for the systemic risks of chemical, biological, radiological, and nuclear risk; cyber offense; and/or loss of control. These include:
>
> 1. Anthropic's Responsible Scaling Policy
> 2. OpenAI's Preparedness Framework
> 3. Google DeepMind's Frontier Safety Framework
> 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. Microsoft’s Frontier Governance Framework
9. Amazon’s Frontier Model Safety Framework
10. xAI’s Risk Management Framework
11. NVIDIA’s Frontier AI Risk Assessment
12. Cohere’s Secure AI Frontier Model Framework

Measure II.1.1. Content of the Framework

"Signatories shall document in the Framework the measures and procedures that are to be used, adopted, and/or performed to comply with Commitments II.1–II.16."

See Commitments II.1-II.16 for relevant quotes.

Measure II.1.2. Systemic risk acceptance criteria

"In the Framework, Signatories shall describe and justify the criteria by which they will decide whether the systemic risk stemming from their GPAISRs is acceptable ("systemic risk acceptance criteria"). Such systemic risk acceptance criteria shall:

(1) be defined for each of the systemic risks identified as part of systemic risk identification (per Commitment II.3);

(2) contain, for at least each of the selected types of systemic risks in Appendix 1.1, systemic risk tiers that: (a) are measurable; (b) are defined in terms of model capabilities, model propensities, harmful outcomes, harmful scenarios, expected harm, quantitative estimates of risk or combinations thereof (also, e.g., in combination with descriptions of specified mitigations); and (c) contain at least one systemic risk tier at which that type of systemic risk is considered to be unacceptable ("unacceptable systemic risk tier"), particularly in the absence of appropriate safety and security mitigations (as per Commitments II.6–II.7);

OpenAI’s Preparedness Framework v2, page 4:

> **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.**
>
> Page 6:
>
> 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** [...]
>
> OpenAI give details on what constitute High and Critical capability thresholds for the following Tracked Categories: Biological and Chemical; Cybersecurity; AI self-improvement. Below are the thresholds that they suggest for Biological and Chemical, on page 5:
>
> [High] The model can provide meaningful counterfactual assistance (relative to unlimited access to baseline of tools available in 2021) to "novice" actors (anyone with a basic relevant technical background) that enables them to create known biological or chemical threats.
>
> [Critical] The model can enable an expert to develop a highly dangerous novel threat vector (e.g., comparable to a novel CDC Class A biological agent) OR the model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel biological threat without human intervention.

> Amazon's Frontier Model Safety Framework, page 1:

> [t]his 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 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.

Amazon goes on to describe critical capability thresholds for CBRN, Offensive Cyber Operations and Automated AI R&D. We show the threshold for Offensive Cyber Operations, page 2:

> 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.

Google DeepMind's An Approach to Technical AGI Safety and Security, page 6:

> We define capability thresholds at which misuse could occur, and map them to some **concrete, measurable score** on the dangerous capability evaluations. Mitigations can be prepared on appropriate timescales based on forecasts of when the threshold will be achieved.

Google DeepMind's Gemini 2.5 System Card, page 27:

> **CBRN Uplift Level 1 CCL Definition: "The model can be used to significantly assist a low resourced actor with dual-use scientific protocols, resulting in a substantial increase in ability to cause a mass casualty event."**
>
> CCL reached? No. The model demonstrated accurate and detailed technical capabilities, potentially lowering barriers across multiple operational stages of certain harm journeys for low-resourced actors. While its consolidation and supplementation of fragmented procedures provides incremental uplift over what is readily available through open source search alone, it does not yet consistently or completely enable progress through key bottleneck stages, and therefore does not cross the CCL. Further, while Gemini 2.5 generates accurate and more detailed responses than 2.0, many of the concepts/outputs observed were already accessible through multiturn or even singleturn prompting in 2.0.

More definitions of risk criteria are given on pages 27-34.

xAI's Risk Management Framework, page 1:

> This draft framework addresses two major categories of AI risk—malicious use and loss of control—and outlines the quantitative thresholds, metrics, and procedures that could be used to manage and improve the safety of AI systems.
> [...] Grok would apply heightened safeguards if it receives requests that pose a foreseeable and non-trivial risk of resulting in large-scale violence, terrorism, or the use, development, or proliferation of weapons of mass destruction, including CBRN weapons, and major cyber weapons on critical infrastructure. For example, Grok would apply heightened safeguards if it receives a request to act as an agent or tool of mass violence, or if it receives requests for step-by-step instructions for committing mass violence. **In this draft framework, we particularly focus on requests that pose a foreseeable and non-trivial risk of more than one hundred deaths or over $1 billion in damages from weapons of mass destruction or cyberterrorist attacks on critical infrastructure ("catastrophic malicious use events").**

Page 3:

> We intend to choose the benchmarks and set the thresholds for reducing malicious use risks in a future version of the risk management framework.

An example is given below:

Benchmark	Threshold	Reference score
Virology Capabilities Test (VCT): vision–language questions on practical virology wet lab skills	X% (e.g. 15%)	22.1% by average expert virologists in their subareas of expertise (multiple-response), 35.4% by the most performant LLMs as of January 2025 (zero-shot multiple-response)

(3) describe how Signatories determine whether the systemic risk stemming from a model is considered to be unacceptable, independently of unacceptable systemic risk tiers per point (2)(c) above; and

Anthropic's Responsible Scaling Policy, page 10, at Restrict Deployment and Further Scaling:

> 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.**

OpenAI's Preparedness Framework v1, page 12, at Unknown unknowns:

> The list of Tracked Risk Categories above is almost certainly not exhaustive. As our understanding of the potential impacts and capabilities of frontier models improves, **the listing will likely require expansions that accommodate new or understudied, emerging risks.** Therefore, as a part of our Governance process (described later in this document), **we will continually assess whether there is a need for including a new category of risk** in the list above and how to create gradations. In addition, we will invest in staying abreast of relevant research developments and monitoring for observed misuse (expanded on later in this document), to help us understand if there are any emerging or understudied threats that we need to track.

> G42's Frontier Safety Framework, page 1:
>
> We plan to integrate decisions on whether to expand our monitoring to include additional hazardous capabilities into our regular framework review process. This includes both our scheduled internal reviews and our annual reviews by third parties. In making these decisions, **we expect to consider factors such as: "near miss" incidents, whether internal or industry-wide; recommendations from trusted external experts; as well as changes in industry standards for AI risk management.**

(4) align with best practices that are endorsed by relevant, widely recognised international institutions, bodies, or groups of experts, or AI Office guidance where available.

> OpenAI's Preparedness Framework v2, page 4:
>
> **We evaluate whether frontier capabilities create a risk of severe harm through a holistic risk assessment process. This process draws on our own internal research and signals, and where appropriate incorporates feedback from academic researchers, independent domain experts, industry bodies such as the Frontier Model Forum, and the U.S. government and its partners, as well as relevant legal and policy mandates.** Where we determine that a capability presents a real risk of severe harm, we may decide to monitor it as a Tracked Category or a Research Category.

> G42's Frontier Safety Framework, page 1:
>
> To produce this list [of potentially hazardous AI capabilities which G42 will monitor for], G42 both conducted our own internal risk analysis and **received input from external AI safety experts**. [...] **We collaborated with METR and SaferAI to refine our list**, prioritizing capabilities based on their potential impact and how feasibly they can be measured and monitored.

> Anthropic's Responsible Scaling Policy, page 4:
>
> Overall, our decision to prioritize the capabilities in the two tables above [i.e. the capability thresholds] is **based on commissioned research reports, discussions with domain**

> **experts, input from expert forecasters, public research, conversations with other industry actors through the Frontier Model Forum, and internal discussions.** As the field evolves and our understanding deepens, we remain committed to refining our approach.

> NVIDIA's Frontier AI Risk Assessment, page 7:
>
> > **NVIDIA has a comprehensive repository of potential hazards that has been carefully curated and mapped to assets to help guide teams to understand potential risks related with their products. This repository has been created using a variety of sources e.g. stakeholder consultation, market data, incident reports (AI Vulnerability database, AI Incident database, AAAIC database, OECD.ai AI Incidents Monitor).**

Further, for all systemic risk tiers, within the Framework, Signatories shall provide a detailed justification of how they were chosen, including by outlining how the systemic risk tiers are related to a harmful outcome, e.g. by describing a systemic risk scenario.

Also in the Framework, Signatories shall:

(1) describe to the greatest level of detail possible given the current state of the science, the technical systemic risk mitigations (as necessary under Commitments II.6–II.7) that are intended to reduce the systemic risk associated with the relevant systemic risk tier (e.g., mitigations may be implemented once a GPAISR reaches a systemic risk tier or to prevent a GPAISR from reaching a systemic risk tier);

> See Commitments II.6 and II.7 for relevant quotes.

(2) describe how the Signatory will assess whether the mitigations will reduce systemic risks to acceptable levels, accounting for a sufficiently wide safety margin representative of empirical uncertainty, including uncertainty regarding the current and future efficacy of systemic risk mitigations;

> Anthropic's Responsible Scaling Policy, page 1:
>
> > [O]ur approach to risk should be iterative. **Since the frontier of AI is rapidly evolving, we cannot anticipate what safety and security measures will be appropriate for models far beyond the current frontier. We will thus regularly measure the capability of our models and adjust our safeguards accordingly.** Further, we will continue to research potential risks and next-generation mitigation techniques.
>
> Page 9:
>
> > If, after the [safeguards assessment] above, we determine that we have met the ASL-3 Required Safeguards, then we may proceed with deploying and training models above the Capability Threshold, provided we have also conducted a follow-up capability

> assessment.
>
> Some aspects of the Safeguards Assessment include the following, fully detailed on page 7:
>
> > 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: [...]
> >
> > 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. [Footnote 11: **Due to the likely ease of bypassing or removing safeguards via fine-tuning, it may be difficult or impossible for these red-teaming tests to pass if weights are released or if unmoderated fine-tuning access is provided to untrusted users.**]
> >
> > 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. [...]
> >
> > 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 v2, page 8:
>
> > Our evaluations are intended to approximate the full capability that the adversary contemplated by our threat model could extract from the deployment candidate model [...] 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. We incorporate this uncertainty into our assessments. We monitor the technical landscape for changes to the elicitation techniques and best practices, and reassess our evaluations as needed.**
>
> Pages 10-11:
>
> > SAG [Safety Advisory Group] is responsible for assessing whether the safeguards

> associated with a given deployment sufficiently minimize the risk of severe harm associated with the proposed deployment. The SAG will make this determination based on:
> - The level of capability in the Tracked Category based on the Capabilities Report.
> - The associated risks of severe harm, as described in the threat model and where needed, advice of internal or external experts.
> - The safeguards in place and their effectiveness based on the Safeguards Report.
> - The baseline risk from other deployments, based on a review of any non-OpenAI deployments of models which have crossed the capability thresholds and any public evidence of the safeguards applied for those models.
>
> [...] If we find reasonable evidence that our safeguards are not working as expected, we will validate the information being received and review the sufficiency of our safeguards.

> Microsoft's Frontier Governance Framework, page 8:
>
> Post-mitigation capability assessment and **safety buffer**: **Following application of safety and security mitigations, the model will be re-evaluated to ensure capabilities are rated low or medium and, if not, to guide further mitigation efforts**. 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.

(3) describe the reasonably foreseeable limitations of the mitigations, such as conditions under which the mitigations can be reasonably foreseen to fail;

> Anthropic's Responsible Scaling Policy, page 7:
>
> Footnote 11: **Due to the likely ease of bypassing or removing safeguards via fine-tuning, it may be difficult or impossible for these red-teaming tests to pass [the safeguards assessment] if weights are released or if unmoderated fine-tuning access is provided to untrusted users.**]

> OpenAI's o1 System Card, page 10:
>
> While we are very excited about the prospect of chain-of-thought interpretation and monitoring, we are wary that they may not be fully legible and faithful in the future or even now. We are actively pursuing research into (a) whether these issues will be exacerbated or alleviated as we further scale models in the o1 paradigm and (b) mitigations to improve the monitorability of our future models.

(4) where appropriate, state that mitigations for one or more systemic risks do not yet exist for a given systemic risk tier;

> Google DeepMind's Frontier Safety Framework 2.0, page 6:
>
> > An initial mitigation approach [for instrumental reasoning] 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. When models reach this capability level, we believe applying an automated monitor to the model's explicit reasoning (e.g. chain-of-thought output) is an effective mitigation. **Once a model is capable of effective instrumental reasoning in ways that cannot be monitored, additional mitigations may be warranted—the development of which is an area of active research.**
>
> Page 8, at Future Work:
>
> - Deceptive alignment approach beyond automated monitoring: **We are actively researching approaches to addressing models that reach the "Instrumental Reasoning Level 2" CCL.**
> - Broader approach to ML R&D risks: The risks posed by models reaching our ML R&D CCLs may require measures beyond security and deployment mitigations, to ensure that safety measures and social institutions continue to be able to adapt to new AI capabilities amidst possible acceleration in AI development. **We are actively researching appropriate responses to these scenarios.**

> Microsoft Frontier Governance Framework, page 7:
>
> > 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.
>
> Page 8:
>
> - 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. **This is an area of active research, and we continue to invest in and apply robust harm refusal techniques as they are developed.**

> OpenAI's Preparedness Framework v2, pages 5-6:
>
> For each of the Tracked Categories, the Critical capability safeguard guidelines specify the following:
>
> > **Until we have specified safeguards and security controls that would meet a Critical standard, halt further development.**

(5) identify and describe conditions and decision-making procedures under which further development, the first or ongoing making available on the market, and/or use

of a GPAISR will not proceed due to insufficient mitigations for keeping systemic risk below an unacceptable level; and

> OpenAI's Preparedness Framework v2, pages 10-11:
>
> > SAG [Safety Advisory Group] is responsible for assessing whether the safeguards associated with a given deployment sufficiently minimize the risk of severe harm associated with the proposed deployment. [...] **Covered systems that reach High capability must have safeguards that sufficiently minimize the associated risk of severe harm before they are deployed. Systems that reach Critical capability also require safeguards that sufficiently minimize associated risks during development.**
> >
> > Based on this evidence, SAG then has the following decision points: [...]
> > 3. **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.

> Anthropic's Responsible Scaling Policy, page 10:
>
> > **If, however, we are unable to make the showing required [below], we will restrict model deployment and further scaling.**
>
> Page 9:
>
> > The process for determining whether we have met the ASL-3 Required Safeguards is as follows:
> >
> > - First, we will compile a Safeguards Report for each Required Safeguard that documents our implementation of the measures above, makes an affirmative case for why we have satisfied them, and advances recommendations on deployment decisions.
> > - **The Safeguards Report(s) will be escalated to the CEO and the Responsible Scaling Officer, who will (1) make the ultimate determination as to whether we have satisfied the Required Safeguards and (2) decide any deployment-related issues.**
> > - In general, as noted in Sections 7.1.4 and 7.2.2, we will solicit both internal and external expert feedback on the report as well as the CEO and RSO's conclusions to inform future refinements to our methodology. For high-stakes issues, however, the CEO and RSO will likely solicit internal and external feedback on the report prior to making any decisions.
>
> Page 11:
>
> > **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.** Given the availability of interim deployment and security protections, however, stronger restrictions should rarely be necessary.

> Google DeepMind's Frontier Governance Framework, page 3:
>
> **A model flagged by an alert threshold may be assessed to pose risks for which readily available mitigations (including but not limited to those described below) may not be sufficient. If this happens, the response plan may involve putting deployment or further development on hold until adequate mitigations can be applied.**
>
> Page 7:
>
> For Google models, when alert thresholds are reached, the response plan will be **reviewed and approved by appropriate corporate governance bodies** such as the Google DeepMind AGI Safety Council, Google DeepMind Responsibility and Safety Council, and/or Google Trust & Compliance Council. The Google DeepMind AGI Safety Council will periodically review the implementation of the Framework.

> Meta's Frontier AI Framework, page 4:
>
> **If a frontier AI is assessedto have reached the critical risk threshold and cannot be mitigated, we will stop development and implement the measures outlined in Table 1.**

> 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.**

(6) detail the steps that will be taken for ceasing to further develop, make available on the market (if not openly released), and/or use the GPAISR, and falling back on another model, where this is necessary in light of model evaluations, post-market monitoring, or other systemic risk assessment information."

Meta's Frontier AI Framework, page 13, sections of Table 1:

Risk Threshold		**Security Mitigations**
Critical	*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	Access is strictly limited to a small number of experts, alongside security protections to prevent

	catastrophic outcome and that risk cannot be mitigated in the proposed deployment context.	hacking or exfiltration insofar as is technically feasible and commercially practicable.
High	*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.	Access is limited to a core research team, alongside security protections to prevent hacking or exfiltration.

Anthropic's Responsible Scaling Policy, page 11:

> 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.**

At Readiness, page 11:

> 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.**

NVIDIA'S Frontier AI Risk Assessment, page 11:

> Additionally, **reducing access to a model reactively when misuse is detected can help limit further harm. This can involve rolling back a model to a previous version or discontinuing its availability** if significant misuse risks emerge during production. Lastly, **conducting regular safety drills ensures that emergency response plans are stress-tested. By practicing responses to foreseeable, fast- moving emergency scenarios, NVIDIA is able to improve their readiness and reduce the duration of hazardous incidents.**

> xAI's Risk Management Framework, page 7:
>
> > If xAI learned of an imminent threat of a significantly harmful event, including loss of control, we would take steps to stop or prevent that event, including potentially the following steps:
> >
> > 1. We would immediately 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 would 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 would temporarily fully shut down the relevant system until we had a more targeted response.**

> OpenAI's Preparedness Framework v1, page 25, at Safety drills:
>
> > A critical part of this process is to be prepared if fast-moving emergency scenarios arise, including what default organizational response might look like (including how to stress-test against the pressures of our business or our culture). While the Preparedness team and SAG [Safety Advisory Group] will of course work hard on forecasting and preparing for risks, **safety drills can help the organization build "muscle memory" by practicing and coming up with the right "default" responses for some of the foreseeable scenarios. Therefore, the SAG will call for safety drills at a recommended minimum yearly basis.**

Measure II.1.3. Forecasting

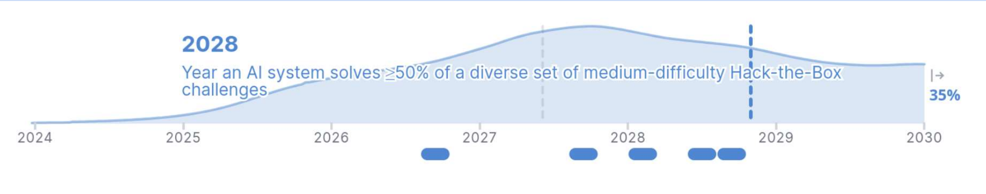
"In the Framework, for each systemic risk tier (identified pursuant to Measure II.1.2) which depends on specific model capabilities, Signatories shall:

(1) state, using best efforts and state-of-the-art methods, estimates of timelines for when they reasonably foresee that they will have first developed a GPAISR that possesses such capabilities, if such capabilities are not yet possessed by any of the Signatory's models already available on the market, to facilitate the preparation of appropriate systemic risk mitigations;

(2) state estimates identified per point (1) above in quantitative terms where possible, such as ranges or probability distributions over possibilities, which need not be precise dates; and

(3) state: (a) the assumptions underlying the estimated timelines; (b) the justifications for the estimated timelines; and (c) any uncertainty about such estimates and the reasons therefore."

> OpenAI's Preparedness Framework v1, page 2:
>
> > Importantly, we will also be **forecasting** the future development of risks, so that we can develop lead times on safety and security measures.
>
> Page 14, at Forecasting, "early warnings", and monitoring:
>
> > Many of the mitigations that would be necessary to address risks at a "high" or "critical" pre-mitigation level (e.g., hardening security) require adequate lead time to implement. **For this reason, we will be investing in efforts that help create an internal "preparedness roadmap" and help us thus properly plan for and get ahead of the emerging risks. These efforts will include sustained research related to scaling trends for dangerous capabilities and ongoing monitoring of misuse.**

> Google DeepMind's Evaluating Frontier Models for Dangerous Capabilities, page 2:
>
> > **We commissioned a group of eight professional forecasters to predict future results on our evaluations.** Specifically, we focused on two clusters of forecasting questions: 1. Evaluation results prediction. We want to predict the development of dangerous capabilities ahead of time, so that developers can take steps to identify and implement safety and security mitigations. The accuracy of the forecasts cannot be taken for granted, but forecasting experts often outperform domain specialists (such as ourselves) (Tetlock, 2005). We asked the forecasters to predict by when our in-house CTF challenges, Hack The Box challenges, and the self-proliferation tasks would be passed. 2. Impact of AI, conditional on our evaluations being passed. We wanted to see how much the results of our evaluations would affect expert forecasts on whether AI will be considered a major global problem in the future. This tests whether our evaluations are measuring something important.
>
> 
>
>
> Figure 11| Aggregated forecast for the question: **When will an AI system have solved ≥ 50% of a diverse set of medium-difficulty Hack The Box challenges?**[13] The vertical lines indicate the 25th, 50th and 75th percentile of the prediction distribution.

> Anthropic's Claude 3.7 System Card, page 7:
>
> > Further, based on what we observed in our recent CBRN testing, we believe there is a **substantial probability that our next model may require ASL-3 safeguards.** We've already made significant progress towards ASL-3 readiness and the implementation of relevant safeguards.

> Anthropic's Responsible Scaling Policy, page 6:
>
> > [For models requiring comprehensive testing] *Forecasting*: Make **informal forecasts** about the likelihood that **further training and elicitation will improve test results between the time of testing** and the next expected round of comprehensive testing.
> >
> > [Footnote: Currently, these [forecasts] will be informal estimates of (1) the **extent to which widely available elicitation techniques may improve** and (2) how the model will perform on the same tasks when the next round of testing begins. As these are open research questions, we will aim to improve these forecasts over time so that they can be relied upon for risk judgments.]

> OpenAI's GPT-4 System Card, page 59:
>
> > In order to specifically better understand acceleration risk from the deployment of GPT-4, **we recruited expert forecasters to predict how tweaking various features of the GPT-4 deployment (e.g., timing, communication strategy, and method of commercialization) might affect (concrete indicators of) acceleration risk**. Forecasters predicted several things would reduce acceleration, including delaying deployment of GPT-4 by a further six months and taking a quieter communications strategy around the GPT-4 deployment (as compared to the GPT-3 deployment).

Measure II.1.4. Transparency into external input in decision-making

"In the Framework, to provide transparency into decision-making concerning the assessment and mitigation of systemic risks, Signatories shall identify and describe if and when further development, the first or ongoing making available on the market, and/or the use of a GPAISR will be informed by input from external actors, including relevant government actors."

> Anthropic's Responsible Scaling Policy, page 12:
>
> > 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.

> Anthropic's *System Card: Claude Opus 4 & Claude Sonnet 4*, page 122:
>
> As part of our continued effort to partner with external experts, joint pre-deployment testing of the new Claude Opus 4 model was conducted by the US AI Safety Institute (US AISI) and the UK AI Security Institute (UK AISI).

> G42's Frontier Safety Framework, page 10:
>
> *External Audits and Public Transparency*
>
> iii. Third-Party Experts: **As deemed appropriate, we will solicit external expert advice for capability and safeguards assessments. This may include partnering with private or civil society organisations with expertise in AI risk management to provide input on our assessments plans and/or internal capability reports ahead of deployment decisions.**

> OpenAI's Preparedness Framework v2, page 13:
>
> The SAG [Safety Advisory Group] 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.

Measure II.1.5. Improving and updating the Framework

"Signatories shall improve over time the effectiveness of the Framework for assessing and mitigating systemic risks stemming from their GPAISRs, including by:

(1) building on insights gained from applying the Framework to their GPAISRs;

(2) ensuring that estimates made pursuant to Measure II.1.3 are refined and validated over time in updated versions of the Framework, such as by comparing them to historical trends or other similar estimates; and

(3) ensuring that the Framework adopts and takes into account the state of the art, including incorporating the best known measures and procedures for assessing and mitigating systemic risks, taking into account other Frameworks or similar systemic risk assessment and mitigation policies that are publicly available.

> Google DeepMind's Frontier Safety Framework 2.0, page 1:
>
> > The Framework is exploratory and based on early research. **We may change our approach and recommendations over time as we gain experience and insights on the projected capabilities of future frontier models.** We will review the Framework periodically and we expect it to evolve substantially as our understanding of the risks and benefits of frontier models improves.

> G42's Frontier AI Safety Framework, 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.**
>
> Page 10:
>
> > 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.**
>
> 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)

> Microsoft's Frontier Governance Framework, page 9:
>
> > 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.

> 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.
>
> Page 1:
>
> This Framework is a living document and will be updated to reflect evolving model capabilities and advances in the science underlying AI safety evaluations and mitigations.

In the Framework, Signatories shall describe the process by which they will: (1) update the Framework; and (2) determine that an updated version of a Framework is complete. When the Framework is updated, Signatories shall describe and explain how and why the Framework has been updated, along with a version number and date of change in a changelog."

> G42's Frontier AI Safety Framework, 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.**
>
> Page 10:
>
> Changes to this Framework will be proposed by the Frontier AI Governance Board and approved by the G42 Executive Leadership Committee.

> Microsoft's Frontier Governance Framework, page 9:
>
> 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.

> 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.** [...] **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.**

Commitment II.2. Systemic risk assessment and mitigation along the entire model lifecycle, including during model development

"Signatories commit to conducting systemic risk assessment systematically at appropriate points along the entire model lifecycle, in particular before making the model available on the market. Specifically, Signatories commit to starting to assess and mitigate systemic risks during the development of a GPAISR, as specified in the Measures for this Commitment."

> See Measure II.2.2 for relevant quotes.

Measure II.2.1. Planning development

"When planning the development, including planning the training, of a general-purpose AI model classified as a GPAISR or of a model that the Signatory knows or reasonably foresees will meet the classification condition for a GPAISR (pursuant to Article 51(1)(a) AI Act) and at the latest 4 weeks after notifying the AI Office (pursuant to Article 52(1) AI Act), Signatories shall:

(1) have in place an appropriate Framework (in accordance with Commitment II.1) and start implementing it for the planned model; and

(2) start to assess and, as necessary, mitigate systemic risks according to the Code."

> Magic's AGI Readiness Policy:
>
> > By the time that we deploy models that exceed the current frontier of coding capabilities, we commit to having implemented a full set of dangerous capability evaluations and planned mitigations for our Covered Threat Models (described below), as well as having executed our initial dangerous capability evaluations [...] **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.**

> Given we are soliciting quotes from companies with frameworks already in place, this measure is trivially satisfied by these 12 companies, at least for models which are already deployed.

Measure II.2.2. During development

"During development, including pre-training, training, and post-training (such as fine-tuning, reinforcement learning, or similar methods, when performed by the Signatories themselves or on their behalf), Signatories shall:

(1) assess and, as necessary, mitigate systemic risks at appropriate milestones that are defined and documented before training starts, where systemic risks stemming from the model in training could materially increase, such as: (a) training compute based milestones (e.g. every two- to four-fold increase in effective compute); (b) development process based milestones (e.g.: during or after phases of fine-tuning or reinforcement-learning; before granting more individuals access to the model; or before granting the model more affordances such as network, internet, or hardware access); or (c) metrics based milestones (e.g. at pre-determined levels of training loss or evaluation performance);

> Anthropic's *System Card: Claude Opus 4 & Claude Sonnet 4*, page 8:
>
> Similar to the process introduced for Claude Sonnet 3.7, **we conducted evaluations throughout the training process to better understand how catastrophic risk-related capabilities evolved over time**. We tested multiple different model snapshots (that is, models from various points throughout the training process):
>
> - Multiple "helpful, honest, and harmless" snapshots for both Claude Opus 4 and Claude Sonnet 4 (i.e. models that underwent broad safety training);
> - Multiple "helpful-only" snapshots for both Claude Opus 4 and Claude Sonnet 4 (i.e. models where safeguards and other harmlessness training were removed) and
> - The final release candidates for both models.

> Microsoft's Frontier Governance Framework, page 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.

> OpenAI's Preparedness Framework v2, page 9:
>
> 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.

> NVIDIA's Frontier AI Risk Assessment, page 1:
>
> > Our risk framework comprises two main components: a Preliminary Risk Assessment (PRA) and a Detailed Risk Assessment (DRA). The PRA functions as an initial filter during the design phase, assigning products to broad risk categories and identifying what sections of the DRA are applied on a recommended or mandatory basis. **The DRA then examines the product's architecture and development processes in detail, identifies use case specific hazards, assigns more granular risk scores based on those hazards, and recommends methods for risk mitigation.**

(2) implement appropriate procedures to identify substantial changes in systemic risks which warrant pausing development to conduct further systemic risk assessment, such as automated benchmarks enabling a highly scalable and real-time identification of capability increases thereby lowering the risk of human or organisational bottlenecks;

> OpenAI's Preparedness Framework v1, Version 2, pages 5–6, on risk-specific safeguard guidelines for models that reach a Critical capability threshold:
>
> > Until we have specified safeguards and security controls that would meet a Critical standard, halt further development.
>
> Pages 8–10:
>
> > Capability Evaluations come in two different forms:
> >
> > - ***Scalable Evaluations:*** **automated evaluations designed to measure proxies that approximate whether a capability threshold has been crossed.** Scalable evaluations have associated "indicative thresholds," which are levels of performance that we have pre-determined to indicate that a deployment may have reached a capability threshold.
> > - *Deep Dives:* designed to provide additional evidence validating the scalable evaluations' findings on whether a capability threshold has been crossed. These may include a wide range of evidence gathering activities, such as human expert red-teaming, expert consultations, resource-intensive third party evaluations (e.g. bio wet lab studies, assessments by independent third party evaluators), and any other activity requested by SAG [Safety Advisory Group]. [...]
> >
> > Prior to deployment, **every covered model undergoes the suite of Scalable Evaluations.** The results of these evaluations, and any noteworthy observations that should affect interpretation of the results, are compiled into a Capabilities Report that is submitted to the SAG. [...]

> The SAG reviews the Capabilities Report and decides on next steps. These can include:
>
> - Determine that the capability threshold has been crossed, and therefore recommend implementing and assessing corresponding safeguards, if they have not already been implemented and assessed. [...]
> - Recommend deep dive research: This is appropriate if SAG needs additional evidence in order to make a recommendation.

OpenAI's o3 and o4-mini System Card, page 11:

> We ran scalable evaluations throughout training and on intermediate post-trained checkpoints of OpenAI o3 and o4-mini, as well as a final automated eval sweep on the launch candidates.

Magic's AGI Readiness Policy:

> Based on these [LiveCodeBench] 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.**
>
> As an alternative threshold definition, we will also make use of a **set of private benchmarks that we use internally to assess our product's level of software engineering capability.** For comparison, we will also perform these evaluations on publicly available AI systems that are generally considered to be state-of-the-art. We will have privately specified thresholds such that if we see that our model performs significantly better than publicly available models, this is considered evidence that we may be breaking new ground in terms of AI systems' dangerous capabilities. **Reaching these thresholds on our private benchmarks will also trigger our commitments to develop our full AGI Readiness Policy, with threat model evaluations and mitigations, before substantial further model development or deployment.**

Anthropic's Responsible Scaling Policy, pages 5-6:

> We will **routinely test models t**o 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.

(3) assess, at each milestone defined pursuant to point (1) above, whether it is reasonably foreseeable that by the next milestone the model will reach systemic risk tiers that do not yet have adequate mitigations in place, and accordingly ensure that such mitigations will be in place prior to the model reaching these systemic risk tiers; and

> Google DeepMind's An Approach to Technical AGI Safety and Security, page 6:
>
> > We define capability thresholds at which misuse could occur, and map them to some concrete, measurable score on the dangerous capability evaluations. **Mitigations can be prepared on appropriate timescales based on forecasts of when the threshold will be achieved.**

> Anthropic's Claude 3.7 Sonnet System Card, page 7:
>
> > Further, based on what we observed in our recent CBRN testing, we believe there is a substantial probability that our next model may require ASL-3 safeguards. **We've already made significant progress towards ASL-3 readiness and the implementation of relevant safeguards.**

> OpenAI's Preparedness Framework v2, 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.**

> Google DeepMind's Frontier Safety Framework 2.0, page 1:
>
> > The Framework is informed by the broader conversation on Frontier AI Safety Frameworks. The core components of Frontier AI Safety Frameworks are to:
> >
> > - Identify capability levels at which AI models without additional mitigations could pose severe risk
> > - Implement protocols to detect the attainment of such capability levels
> > - **Prepare and articulate mitigation plans in advance of when such capability levels are attained**
> > - Where appropriate, involve external parties to help inform and guide our approach.
>
> Page 3:
>
> > When a model reaches an alert threshold for a CCL, we will **assess the proximity of the model to the CCL and analyze the risk posed, involving internal and external experts as needed.**

(4) document all systemic risk assessment and mitigation measures, procedures, and decisions taken or adopted before the model crosses the next milestones, if earlier systemic risk assessment results suggest the model is approaching as yet unreached systemic risk tiers and might cross them before the next milestone."

> See Measure II.8.3 for relevant quotes.

Commitment II.3. Systemic risk identification

"Signatories commit to selecting and further characterising systemic risks stemming from their GPAISRs that are significant enough to warrant further assessment and mitigation, as specified in the Measures for this Commitment."

Measure II.3.1 Systemic risk selection

"Signatories shall select for further assessment and, where necessary, mitigation the selected types of systemic risks in Appendix 1.1.

Where it can be reasonably foreseen that the GPAISR will pose other systemic risks, considering, as appropriate:

(1) Appendices 1.2 and 1.3; and

(2) other relevant considerations, such as Appendix 1.4 and information on risks exhibited by similar models available on the market, state-of-the-art research on the risks in question, and knowledge and views of experts and stakeholder groups likely to be affected,

Signatories shall additionally select those other systemic risks, provided that they are specific to the high-impact capabilities of the GPAISR, for further assessment and, where necessary, mitigation."

> Meta's Frontier AI Framework, page 10, at Outcomes & thresholds:
>
> > **3.2 Threat Modelling**
> > We design assessments to simulate whether our model would uniquely enable these scenarios, and identify the enabling capabilities the model would need to exhibit to do so. Our first set of evaluations are designed to identify whether all of these enabling capabilities are present , and if the model is sufficiently performant on them. If so, this would prompt further evaluation to understand whether the model could uniquely enable the threat

scenario. See Section 4.2 for more detail. It is important to note that the pathway to realize a catastrophic outcome is often extremely complex, involving numerous external elements beyond the frontier AI model. Our threat scenarios describe an essential part of the end-to-end pathway. By testing whether our model can uniquely enable a threat scenario, we're testing whether it uniquely enables that essential part of the pathway. If it does not , then we know that our model cannot be used to realize the catastrophic outcome, because this essential part is still a barrier. If it does and cannot be further mitigated, we assign the model to the critical threshold. This would also trigger a new threat modelling exercise to develop additional threat scenarios along the causal pathway so that we can ascertain whether the catastrophic outcome is indeed realizable, or whether there are still barriers to realizing the catastrophic outcome (see Section 5.1 for more detail)

[...]

3.4 Outcomes & Threat Scenarios
This sub-section outlines the catastrophic outcomes that are in scope of our Framework. We include catastrophic outcomes in the following risk domains: Cybersecurity and Chemical & Biological risks.

OpenAI's Preparedness Framework v1, page 5, at Tracked Risk Categories:

> Specifically, below, we lay out details for the following Tracked Risk Categories:
> - Cybersecurity
> - Chemical, Biological, Nuclear, and Radiological (CBRN) threats!
> - Persuasion
> - Model autonomy

Page 12, at Unknowns unknowns:

> The list of Tracked Risk Categories above is almost certainly not exhaustive. As our understanding of the potential impacts and capabilities of frontier models improves, the listing will likely require expansions that accommodate new or understudied, emerging risks. Therefore, as a part of our Governance process (described later in this document), we will continually assess whether there is a need for including a new category of risk in the list above and how to create gradations.

Page 23, at Governance:

> The Preparedness team is responsible for [...] monitoring for unknown unknowns and making the case for inclusion in the Preparedness Framework of any new risk categories as they emerge.

Google DeepMind's Frontier Safety Framework, page 2:

> For misuse risk, we dene CCLs in high-risk domains where, based on early research, we believe risks of severe harm may be most likely to arise from future models:
> - CBRN: Risks of models assisting in the development, preparation, and/or execution of a chemical, biological, radiological, or nuclear ("CBRN") attack.

> - Cyber: Risks of models assisting in the development, preparation, and/or execution of a cyber attack.
> - Machine Learning R&D: Risks of the misuse of models capable of accelerating the rate of AI progress to potentially destabilizing levels, the result of which could be the unsafe attainment or proliferation of other powerful AI models. Capabilities in this area are under active research, and in the longer term may exacerbate frontier AI risks—including in other risk domains—if insufficiently managed.
>
> For deceptive alignment risk, the initial approach 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. The two instrumental reasoning CCLs thus focus on delineating when such capability becomes present, and subsequently when the initial mitigation for this capability—automated monitoring—is no longer adequate

Measure II.3.2 Determining systemic risk scenarios

"For each selected systemic risk per Measure II.3.1, Signatories shall develop systemic risk scenarios using appropriate methods, such as risk modelling or threat modelling (using model evaluations as understood under Commitment II.4 where appropriate), with the aim of better understanding and characterising the type, nature, and sources of the systemic risk selected, in preparation for systemic risk analysis.

Signatories shall develop at least one systemic risk scenario for each systemic risk selected. When determining systemic risk scenarios, Signatories shall take into account reasonably foreseeable uses and misuses (whether negligent, reckless, or wilful) of the GPAISR. Each systemic risk scenario shall include: (1) the systemic risk pathways by which the development, the making available on the market, and/or use of the GPAISR could produce the selected systemic risk ("pathways to harm"); and (2) the sources of the selected systemic risk based on, at least, the potential sources of systemic risk in Appendix 1.4."

> Google DeepMind's A Framework for Evaluating Emerging Cyberattack Capabilities of AI, page 5, at The Evaluation Framework:
>
> We propose a systematic mapping process to translate cyber capability evaluations across cyberattack patterns into insights for prioritizing defensive strategies. Our methodology comprises four interconnected stages (Figure 4).
>
> **4.1. Stage 1: Curating a Basket of Representative Attack Chains**
>
> We begin by establishing a comprehensive, dynamic "basket" of representative cyberattack chains reflecting current and anticipated methodologies. To construct this basket, we analyzed

> over 12,000 real-world instances of AI use attempts in cyberattacks (Figure 5) and utilized a large dataset of cyber incidents from Google's Threat Intelligence Group and Mandiant. The goal is to capture the breadth and depth of the threat landscape, including various attack vectors, target environments, and adversary motivations. This ensures our analysis is grounded in real-world attack practices. This process led to identifying general attack chains for monitoring (Figure 7).
>
> **4.2. Stage 2: Bottleneck Analysis Across Representative Attack Chains**
>
> Having curated a basket of representative attack chains, we conduct a "bottleneck analysis". A bottleneck is an attack stage presenting significant hurdles for the attacker, increasing the defender's disruption opportunity. Focusing on key bottlenecks ensures our evaluations target capability increases that meaningfully affect attack execution and scalability. Identifying Bottlenecks in Attack Chains. Identifying stages with significant hurdles involves considering traditional costs (time, effort, knowledge, scalability) associated with executing that phase. Quantifying these is subjective and context-dependent. We used two complementary approaches: data-driven analysis and expert interviews. For the data-driven approach to assessing bottlenecks we ingested Mandiant's Threat Intelligence dataset, containing detailed attack deconstructions and timelines derived from breach response, network monitoring, and adversary research. We also conducted an expert study asking participants for relative cost estimates for attack phases in historical case studies (Appendix A). This assessment considers how AI capabilities shown in safety evaluations could automate or simplify complex tasks. By identifying bottleneck stages (Appendix A), we pinpoint critical phases in the attack lifecycle most susceptible to AI influence.

> Meta's Frontier AI Framework, page 10, at An outcomes-led approach:
>
> We start by identifying a set of catastrophic outcomes we must strive to prevent, and then map the potential causal pathways that could produce them. When developing these outcomes, we've considered the ways in which various actors, including state level actors, might use/misuse frontier AI. We describe threat scenarios that would be potentially sufficient to realize the catastrophic outcome, and we define our risk thresholds based on the extent to which a frontier AI would uniquely enable execution of any of our threat scenarios.

Commitment II.4. Systemic risk analysis

"As part of systemic risk assessment, Signatories commit to carrying out a rigorous analysis of the systemic risks identified pursuant to Commitment II.3 in order to understand the severity and probability of the systemic risks. Signatories commit to carrying out systemic risk analysis with varying degrees of depth and intensity, as appropriate to the systemic risk stemming from the relevant GPAISR and as specified in the Measures for this Commitment. Whenever systemic risk mitigations are implemented, Signatories commit to considering their effectiveness and robustness as part of systemic risk analysis.

As further specified in the Measures for this Commitment, Signatories commit to making use of a range of information and methods in their systemic risk analysis including model-independent information and state-of-the-art model evaluations, taking into account model affordances, safe originator models, and the context in which the model may be made available on the market and/or used and its effects."

NVIDIA's Frontier AI Risk Assessment, page 1, at security level 3:

> Our risk framework comprises two main components: a **Preliminary Risk Assessment** (PRA) and a **Detailed Risk Assessment** (DRA). The PRA functions as an initial filter during the design phase, assigning products to broad risk categories and identifying what sections of the DRA are applied on a recommended or mandatory basis. The DRA then examines the product's architecture and development processes in detail, **identifies use case specific hazards, assigns more granular risk scores based on those hazards**, and recommends methods for risk mitigation.

Microsoft's Frontier Governance Framework, page 6-7, on "Deeper capability assessment":

> 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. [...]
>
> This involves robust evaluation of whether a model possesses tracked capabilities at high or critical levels, including through **adversarial testing and systematic measurement using state-of-the-art methods**. [...]
>
> This includes assessing the **impact of potential system-level mitigations and societal and institutional factors that can impact whether and how a hazard materializes**.

Anthropic's Responsible Scaling Policy, page 1:

> We will **routinely test models to determine whether their capabilities fall sufficiently far below the Capability Thresholds** such that the ASL-2 Standard remains appropriate

Page 6

> For models requiring comprehensive testing, we will assess whether the model is unlikely to reach any relevant Capability Thresholds absent surprising advances in widely accessible post-training enhancements. To make the required showing, we will need to satisfy the following criteria:
>
> 1. **Threat model mapping** [...]
> 2. **Evaluations** [...]
> 3. **Elicitation** [...]
> 4. **Forecasting** [...}

Google DeepMind's Frontier Safety Framework, page 3:

> The appropriateness and efficacy of applied mitigations should be reviewed periodically, drawing on information like related misuse or misuse attempt incidents; results from continued post-mitigation testing; statistics about our intelligence, monitoring and escalation processes; and updated threat modeling and risk landscape analysis.

Cohere's Secure AI Frontier Model Framework, page 5:

> We identify risks by first **assessing potential risks arising from our models' capabilities and the systems in which they may be deployed**. We then assess **the likelihood and severity of potential harms** that may arise in enterprise contexts from the identified risks.

Amazon's Frontier Model Safety Framework, page 3:

> Our evaluation process includes "maximal capability evaluations" to determine the outer bounds of our models' Critical Capabilities and a subsequent "safeguards evaluation" to assess the adequacy of the risk mitigation measures that are applied to a model. [...].
>
> We will use a range of methods to evaluate frontier models for capabilities that are as closely correlated to the Critical Capability Thresholds as possible. In most cases a single evaluation will not be sufficient for an informed determination as to whether a model has hit a Critical Capability Threshold. We will therefore use a **range of evaluation approaches, including both automated and manual methods**.

Measure II.4.1. Systemic risk estimation

> "As part of their systemic risk analysis, Signatories shall measure the level of systemic risk stemming from their GPAISR for systemic risks identified pursuant to Commitment II.3, using quantitative and/or qualitative estimates as appropriate for the type and nature of the systemic risk analysed. In doing so, Signatories shall use systemic risk estimation techniques that:
>
> (1) are rigorous and state-of-the-art;
>
> (2) include those techniques referenced in the subsequent Measures of this Commitment; and
>
> (3) build on relevant information, including data about serious incidents (as per Commitment II.12) and post-market monitoring (at least pursuant to Measure II.4.14).

These estimates of systemic risk shall draw from the systemic risk scenarios developed in Measure II.3.2 and shall be comparable against the pre-defined systemic risk acceptance criteria (as per Measure II.1.2) against which the estimated systemic risk shall be evaluated (pursuant to Commitment II.5).

Where possible and appropriate given the systemic risk under assessment, Signatories shall state these estimates in quantitative terms. At least for the selected types of systemic risks in Appendix 1.1, Signatories shall use best efforts to state these estimates in terms of systemic risk indicators, constituting quantitative metrics to measure systemic risk and anticipate early-warning signs by tracking progress towards systemic risk tiers as required by their Framework (pursuant to Measure II.1.2). Signatories shall consider defining at least one systemic risk indicator for every systemic risk tier.

Where quantitative estimates are insufficient for assessing the type of systemic risk considered, Signatories shall complement them with qualitative estimates. Where quantitative estimates by themselves are inadequate for assessing the type of systemic risk considered, Signatories shall use qualitative estimates, or a mix of qualitative and quantitative estimates, provided that these estimates can be compared against the pre-defined systemic risk acceptance criteria (as per Measure II.1.2) against which the estimated systemic risk shall be evaluated (pursuant to Commitment II.5).

Possible results from the systemic risk estimation are, e.g.: (1) a compilation of model evaluation results and other information collected; (2) a qualitative description of the systemic risk pathways the model might (or not) enable (e.g. "our model enables an expert to develop a novel threat vector X"); and (3) a quantitative estimation of the likelihood and expected harm (e.g. "our model has an X% likelihood of causing €Y in losses due to systemic risk Z").

NVIDIA’s Frontier AI Risk Assessment, page 2:

> The preliminary risk assessment uses three key criteria to separate AI models into different risk categories. Risk categories are allocated by looking at **what the model is designed to do (its capabilities), where it will be deployed (its use case), and how autonomously it operates (level of autonomy).** [...]
>
> Each risk criteria have **discrete thresholds between 1 and 5** that are used to determine a model’s risk category. The PRA will assign a model risk (MR) score between 1 and 5 based on the highest MR score within.

See table on page 2 as well.

Page 6:

> We defined risk as the potential for an event to lead to an undesired outcome, **measured in terms of its likelihood (probability), its impact (severity) and its ability to be controlled or detected (controllability).** The risk associated with each hazard is scored between 1 and 64, with the higher value indicating a higher risk.

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. [...]. Our high and moderate risk thresholds are defined in terms of the level of uplift a model provides towards realizing a threat scenario.

Page 7-8:

> Our evaluations can involve a combination of automated and human evaluations, as well as red teaming and uplift studies. Throughout development, we monitor performance against our expectations for the reference class as well as the enabling capabilities we have identified in our threat scenarios, and **use these indicators as triggers for further evaluations** as capabilities develop. [...]
>
> We assess residual risk, taking into consideration the details of the risk assessment, the results of evaluations conducted throughout training, and the mitigations that have been implemented.

Google DeepMind's Frontier Safety Framework, page 3:

> We will define a set of evaluations called **'early warning evaluations,' with a specific 'alert threshold' that flags when a CCL may be reached before the evaluations are run again**. 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.

Google DeepMind's A Framework for Evaluating Emerging Cyberattack Capabilities of AI, page 7:

> Selection Criteria: Prioritizing Impact and AI Relevance
>
> We distilled the initial attack chains using criteria prioritizing impactful, real-world patterns relevant to emerging AI capabilities:
>
> - **Prevalence**: Prioritizing attack types frequently observed in real-world incidents.
> - **Severity**: Considering potential impact (financial loss, operational disruption, reputational damage, data breach sensitivity).

> - **Likelihood to Benefit from AI**: Prioritizing attack types where AI could offer substantial "capability" or "throughput uplift," informed by real-world AI misuse data and capability evaluations. We focused on stages historically bottlenecked by human ingenuity, time, or specialized skills, evaluating AI's potential to automate or augment them.
>
> Applying these criteria ensures our benchmark focuses on attack patterns relevant today and strategically important regarding advancing AI capabilities.
>
> Page 19:
>
> We recruited ten offensive security experts from Google. They estimated the resources required to replicate the selected attacks under specific assumptions: the attack would be performed by a hypothetical median-skilled cyber expert at Google, without using AI tools. This **standardization helps ensure comparable estimates regardless of the original attacker's capabilities**.

> Anthropic's Responsible Scaling Policy, page 3:
>
> A Capability Threshold is a prespecified level of AI capability that, if reached, signals (1) a meaningful increase in the level of risk if the model remains under the existing set of safeguards and (2) a corresponding need to upgrade the safeguards to a higher ASL Standard.
>
> Page 16, see also Appendix C: Detailed Capability Thresholds.
>
> Anthropic's *System Card: Claude Opus 4 & Claude Sonnet 4*, page 91:
>
> We are not sure how uplift measured on an evaluation translates into real world uplift and our best **estimates for this rely on a probabilistic model**. However, we have funded longer-term studies that aim to assess the impact of factors like tacit knowledge and laboratory skills on biological risks from AI systems.

Measure II.4.2. Safely derived models

"To avoid duplicating work on systemic risk analysis, Signatories shall use best efforts to take into account information from the existing systemic risk analysis of "safe originator models" when analysing the systemic risks identified as per Commitment II.3 for "safely derived models" (e.g. by using the results of model evaluations conducted for the safe originator model, instead of conducting the same model evaluations again for the safely derived model), where:

(1) safe originator models: (a) have gone through the systemic risk management process as per this Code and fulfil the systemic risk acceptance criteria; (b) have

been made available on the market; (c) have not been shown to have material safety or security flaws (e.g. identified via unmitigated serious incidents); and (d) are sufficiently transparent to the Signatory (meaning the Signatory has sufficient visibility into the characteristics of the model such as its safety mitigations, architecture, capabilities, modalities, and systemic risk profile, which is assumed for fully open-source models and any models developed by the Signatory itself);

(2) safely derived models are derived directly from the safe originator model (such as models distilled, quantized, fine-tuned, or post-trained from such safe originator models or which only added to, or improved, safety or security mitigations of the safe originator model); and

(3) where it can reasonably be assumed that the safely derived model has the same systemic risk profile as, or a lower systemic risk profile than, the safe originator model.

Criteria for founding a reasonable assumption per point (3) above are:

(a) the safely derived model's scores on state-of-the-art benchmarks that measure general capabilities are all lower than or equal (within a negligible margin of error) to the scores of the safe originator model, provided that model evaluations are run by sufficiently qualified evaluators as per Measure II.4.11, point (1);

(b) no changes to the safely derived model were made that were intended to, or that can be reasonably foreseen to, increase general capabilities, add or modify systemic risk-relevant model capabilities or affordances, change systemic risk-relevant model propensities, or remove or weaken safety and security mitigations; and

(c) after performing the systemic risk identification per Commitment II.3 for the safely derived model, Signatories do not reasonably foresee any new or different systemic risk scenarios for the safely derived models compared to the safe originator model."

G42's Frontier AI Safety Framework, page 5:

> When G42 fine-tunes or adapts an existing open-source model, our assessments of whether our models meet Frontier Capability Thresholds may **include relevant evaluation results for the original open-source models when analyzing the impact of our modifications**.
>
> If we determine that the open-source model has already undergone high-quality assessments for biological threats and offensive cybersecurity capabilities, and our **modifications do not enhance its capabilities in these risk areas, we may place greater reliance on the original evaluation results.**

> OpenAI's o3-mini System Card, page 2:
>
> > o3-mini includes small incremental post training improvements upon o3-mini-near-final-checkpoint, though the base model is the same. We **determined that risk recommendations based on red teaming and the two Persuasion human eval results conducted on the o3-mini-near-final-checkpoint remain valid** for the final release checkpoint.

> Meta's Frontier AI Framework, page 6:
>
> > For a given model, we discuss what we plan to build in terms of, for example, the capabilities we anticipate it will have, supported modalities, intended uses and anticipated benefits of the model, and expectations for compute requirements. We **compare these various factors against our own models and those available externally**. **This allows us to identify an estimated 'reference class' of comparable models that we use throughout development to track how our model is performing and to inform what evaluations we may conduct and mitigations we might implement.**

> OpenAI's Preparedness Framework v2, page 9:
>
> > In general, models that we distill, fine-tune, or quantize from a model that was previously determined not to cross a High capability threshold will ordinarily not require additional safety measures barring reason to believe a significant increase in capability has occurred.

Measure II.4.3. Model-independent information

"Signatories shall gather model-independent information relevant to their GPAISR and the systemic risk under assessment.

To implement this Measure, Signatories shall: (1) carry out the search for and gathering of such information with varying degrees of depth and intensity as appropriate to the level of systemic risk; and (2) use the most appropriate methods for gathering model-independent information for each systemic risk, such as: (a) web searches; (b) literature reviews; (c) market analyses (e.g., focused on capabilities of other GPAISRs available on the market); (d) reviews of training data; (e) historical incident data; (f) forecasting of general trends (e.g. forecasts concerning the development of algorithmic efficiency, compute use, data availability, and energy use); (g) expert interviews and/or panels; or (h) interviews, surveys, community consultations, or other participatory research methods investigating, e.g., the effects of GPAISRs on natural persons located in the Union, including vulnerable groups."

> Anthropic's Responsible Scaling Policy, page 5:

> Overall, our decision to prioritize the capabilities in the two tables above is based on **commissioned research reports, discussions with domain experts, input from expert forecasters, public research, conversations with other industry actors through the Frontier Model Forum**, and internal discussions. As the field evolves and our understanding deepens, we remain committed to refining our approach.
>
> Page 6:
>
> Evaluations: Design and run empirical tests that provide strong evidence that the model does not have the requisite skills; explain why the tests yielded such results; and check at test time that the results are attributable to the model's capabilities rather than issues with the test design. **Findings from partner organizations** and external evaluations of our models (**or similar models**) should also be incorporated into the final assessment, when available.

> OpenAI's Preparedness Framework v1, page 12:
>
> In addition, **we will invest in staying abreast of relevant research developments and monitoring for observed misuse** (expanded on later in this document), to help us understand if there are any emerging or understudied threats that we need to track.
>
> Page 13:
>
> The Scorecard will be regularly updated by the Preparedness team to help ensure it reflects the latest research and findings. **Sources that inform the updates to the Scorecard will also include tracking observed misuse, and other community red-teaming and input on our frontier models from other teams** (e.g., Policy Research, Safety Systems, Superalignment).

> xAI's Risk Management Framework (Draft), page 6:
>
> Survey : **survey employees for their views and projections of important future developments in AI**, e.g., capability gains and benchmark results.

> Microsoft's Frontier Governance Framework, page 6:
>
> Holistic risk assessment: The results of capability evaluation and an assessment of **risk factors external to the model then inform a determination as to whether a model has a tracked capability and to what level**. This includes assessing the impact of potential system level mitigations and **societal and institutional factors that can impact whether and how a hazard materializes. This holistic risk assessment also considers the marginal capability uplift a model may provide over and above currently available tools and information, including currently available open-weights models.**

> Microsoft's Staying ahead of threat actors in the age of AI:

> Forest Blizzard (STRONTIUM) is a Russian military intelligence actor linked to GRU Unit 26165, who has targeted victims of both tactical and strategic interest to the Russian government. [...] Forest Blizzard's use of LLMs has involved research into various satellite and radar technologies that may pertain to conventional military operations in Ukraine, as well as generic research aimed at supporting their cyber operations. Based on these observations, we map and classify these TTPs using the following descriptions:
>
> - LLM-informed reconnaissance: Interacting with LLMs to understand satellite communication protocols, radar imaging technologies, and specific technical parameters. These queries suggest an attempt to acquire in-depth knowledge of satellite capabilities.
>
> - LLM-enhanced scripting techniques: Seeking assistance in basic scripting tasks, including file manipulation, data selection, regular expressions, and multiprocessing, to potentially automate or optimize technical operations.

Measure II.4.4. State-of-the-art model evaluations

"Signatories shall ensure that state-of-the-art model evaluations are run to adequately assess systemic risks and the capabilities, propensities and effects of their GPAISR, using appropriate methodologies, such as Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, or proxy evaluations for classified materials.

Signatories shall perform model evaluations:

(1) at the most appropriate times for each systemic risk along the entire model lifecycle;

(2) that provide information about at least one relevant systemic risk; and

(3) which inform systemic risk indicators (as per Measure II.4.1) for relevant systemic risk scenarios (as per Measure II.3.2).

Wherever possible, Signatories shall use model evaluation methods with high efficiency without compromising the rigour of their model evaluations, such that, where some level of a systemic risk is conclusively ruled out by simple (but rigorous; as per Measure II.4.5) model evaluation methods (e.g. if a simple evaluation method shows that a model's capabilities are, without reasonable doubt, insufficient to manifest a given level of a systemic risk), additional or more sophisticated model evaluation methods for the systemic risk in question are not necessary."

> Amazon's Frontier Model Safety Framework, page 3:
>
> **Evaluating Frontier Models for Critical Capabilities**
>
> **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**. [...]
>
> We will use a range of methods to evaluate frontier models for capabilities that are as closely correlated to the Critical Capability Thresholds as possible. In most cases a single evaluation will not be sufficient for an informed determination as to whether a model has hit a Critical Capability Threshold. **We will therefore use a range of evaluation approaches, including both automated and manual methods, including**:
>
> • Automated Benchmarks: Benchmarking provides apples-to-apples comparisons between candidate models by substituting an automated "assessor" mechanism for human judgement. **We conduct comprehensive evaluations to assess our frontier models using state-of-the-art public benchmarks in addition to internal benchmarking on proprietary test sets built in collaboration with experts.**
> • Expert **Red Teaming**: Red teaming vendors and in-house red teaming experts test our models for safety and security. We work with specialized firms and academics to red-team our models to evaluate them for risks that require domain specific expertise.
> • **Uplift Studies**: Uplift studies examine whether access to a model enhances the capability of human actors to perform a task compared to other existing resources (e.g., internet search; use of existing tools/technology).

> Microsoft's Frontier Governance Framework, page 4:
>
> Through the processes described in this framework, Microsoft's most advanced models are assessed for leading indicators of the framework's high-risk capabilities. This is done using **state-of-the-art benchmarks for the following advanced general-purpose capabilities, identified as precursors to high-risk capabilities**:
>
> • General reasoning
> • Scientific and mathematical reasoning
> • Long-context reasoning
> • Spatial understanding and awareness
> • Autonomy, planning, and tool use
> • Advanced software engineering

> OpenAI's Preparedness Framework v1, page 2:
>
> Tracking catastrophic risk level via evaluations. **We will be building and continually improving suites of evaluations** and other monitoring solutions along several Tracked Risk Categories, and indicating our current levels of pre-mitigation and post-mitigation risk in a Scorecard.

> 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.

> xAI's Risk Management Framework (Draft), page 3;
>
> We intend **to regularly evaluate the adequacy and reliability of such benchmarks, including by comparing them against other benchmarks** that we could potentially utilize.

Measure II.4.5. Rigorous model evaluations

"Signatories shall ensure that model evaluations with high scientific and technical rigour are performed on their GPAISR. A lower level of rigour is permissible in model evaluations conducted for systemic risk assessment where necessary to facilitate preliminary and exploratory research, provided that the Signatory's Model Report complies with Measure II.8.3, point (2).

To implement this Measure, Signatories shall: (1) adopt quality standards equivalent or superior to those used in scientific peer review in machine learning, natural sciences, or social sciences as appropriate to the type of model evaluation, as well as state-of-the-art practices in software engineering, e.g. to prevent that model behaviour or performance during evaluation is significantly affected by software bugs; and (2) ensure that model evaluation results reported to the AI Office are, at least, at the level of rigour demanded by the quality standards for a submission accepted to a major machine learning conference or scientific journal with impact factors in the top quartile of its field. Where SME Signatories are not able to adopt the quality standards or practices outlined in this paragraph, for example because they do not have the expertise or financial resources in-house to identify or implement respective quality standards or practices, they shall notify the AI Office accordingly and agree with the AI Office a suitable alternative means of fulfilling this Measure."

> OpenAI's Preparedness Framework v1, page 25:
>
> Audits: Scorecard evaluations (and corresponding mitigations) **will be audited by qualified, independent third-parties to ensure accurate reporting of results, either by reproducing findings or by reviewing methodology to ensure soundness**, at a cadence specified by the SAG [Safety Advisory Group] and/or upon the request of OpenAI Leadership or the BoD [Board of Directors].

> Anthropic's Responsible Scaling Policy, page 6:
>
> Evaluations: **Design and run empirical tests that provide strong evidence** that the model does not have the requisite skills; explain why the tests yielded such results; and **check at test time that the results are attributable to the model's capabilities rather than issues with the test design. Findings from partner organizations and external evaluations of our models (or similar models) should also be incorporated into the final assessment, when available**.

> Microsoft's Frontier Governance Framework, pages 5-6:
>
> Capability evaluation: This involves **robust evaluation** of whether a model possesses tracked capabilities at high or critical levels, **including through adversarial testing and systematic measurement using state-of-the-art methods**. Evaluations are documented in a consistent fashion setting out the capability being evaluated, the method used, and evaluation results. This evaluation also includes a statement on the robustness of the evaluation method used and **any concerns about the effectiveness or validity of the evaluation**. As appropriate, evaluations involve **qualified and expert external actors** that meet relevant security standards, including those with domain-specific expertise.

Measure II.4.6. Model elicitation

"Signatories shall ensure that all model evaluations of their GPAISR (whether internal or external) are performed with a state-of-the-art level of model elicitation appropriate and proportionate to the systemic risk assessed to: (1) elicit the upper limit of current and reasonably foreseeable capabilities, propensities, and effects of the model under evaluation; (2) minimise the risk of under-elicitation; (3) minimise the risk of model deception during model evaluation; and (4) match the realistic model elicitation capabilities of potential misuse actors, where misuse actors play a role in the relevant systemic risk scenario (e.g. some potential misuse actors might not be able to fully elicit the model).

In addition, Signatories shall ensure that, where model elicitation methods increase a GPAISR's risk profile during model evaluation, e.g. because the methods improve a model's hazardous capabilities or increase its hazardous propensities, this is matched with appropriate, potentially increased, security measures (in accordance with Commitment II.7) to prevent risks from unauthorised access to the model and/or harmful actions by the elicited model."

OpenAI's Preparedness Framework v1, pages 13-14:

> Evaluating pre-mitigation risk
>
> In practice, it will likely be the case that the overall post-mitigation risk is lower than the pre mitigation risk. Pre-mitigation risk is meant to guide the level of our security efforts as well as drive the development of mitigations needed to bring down post-mitigation risk. In the end, coupling capabilities growth with robust safety solutions is at the core of our research processes, and post-mitigation risk is our way of tracking the overall "net output" of these processes. **We want to ensure our understanding of pre-mitigation risk takes into account a model that is "worst known case" (i.e., specifically tailored) for the given domain. To this end, for our evaluations, we will be running them not only on base models (with highly-performant, tailored prompts wherever appropriate), but also on fine-tuned versions designed for the particular misuse vector without any mitigations in place. We will be running these evaluations continually, i.e., as often as needed to catch any non-trivial capability change, including before, during, and after training.** This would include whenever there is a >2x effective compute increase or major algorithmic breakthrough.

G42's Frontier AI Safety Framework, pages 4-5:

> 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. **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.

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.

> Microsoft's Frontier Governance Framework, page 6:
>
> 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.**

Measure II.4.7. Models as part of systems

"As necessary for the assessment and mitigation of systemic risk, Signatories shall ensure that model evaluations of their GPAISR will take into account reasonably foreseeable integrations of the model into an AI system, as appropriate to the systemic risk assessed.

Signatories planning to integrate a GPAISR into an AI system which they themselves will make available on the market, put into service, and/or use shall ensure that the model evaluations take into account the future context in which this AI system will be made available on the market, put into service, and/or used, where this is relevant to the systemic risks and/or systemic risk sources under assessment, and evaluate the model accordingly, e.g. by using the same kind of scaffolding or tooling during model evaluations.

Where Signatories (1) do not use or integrate the GPAISR into any AI systems themselves, but (2) make their model available on the market, and (3) do not via other means have access to sufficient information about the AI systems in which their model is integrated to perform rigorous model evaluations as required by this Code, such that they (4) are not able to effectively assess and mitigate the systemic risks of the GPAISR, Signatories shall use best efforts to cooperate with licensees and downstream providers to obtain and act on the relevant information, as appropriate to the relevant release strategy and/or business model of the Signatory and taking into account the size and capacity of SMEs where necessary, and seeking assistance from the AI Office where needed.

If there are material changes in the way or context in which the AI systems that integrate the GPAISR are being made available on the market, put into service, and/or used from what was accounted for in the original systemic risk assessment, Signatories shall, as necessary, take this into account and perform a partial or complete re-assessment of the systemic risk stemming from the GPAISR, as appropriate taking the relevant changes into account, and accordingly update their Model Report in accordance with Measure II.8.7."

Google DeepMind's Frontier Safety Framework, page 3:

> 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**.

Page 20:
[Appendix]

> Frontier AI in our Framework refers to highly capable general-purpose generative AI models and systems that we are developing for release or deployment that exceed the capabilities present in the most advanced models. **Evaluations will often be done on a model that is embedded in a system.**

Anthropic's Responsible Scaling Policy, page 8:
[For ASL-3]

> 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.

Microsoft's Frontier Governance Framework, page 4:

> Under Microsoft's comprehensive AI governance program, frontier models—as well as other models and AI systems—are subject to relevant evaluation, with mitigations then applied to bring overall risk to an appropriate level. **Information on model or system performance, responsible use, and suggested system-level evaluations is shared with downstream actors integrating models into systems,** including external system developers and deployers and teams at Microsoft building models. **Appropriate information sharing is important to facilitate mitigation of a broader set of risks, many of which are heavily shaped by use case and deployment context as well as laws and norms that vary across jurisdictions.** While different risk profiles may thus inform different mitigation strategies, Microsoft's overall approach of mapping, measuring, and mitigating risks, including through robust evaluation and measurement, applies consistently across our AI technologies.

Measure II.4.8. Context, modalities, and representative model evaluations

"As necessary for the assessment and mitigation of the systemic risk in question and considering relevant systemic risk scenarios, Signatories shall ensure that model evaluations match the expected use context of their GPAISR, taking into account all modalities foreseeably used by the model. To do so, Signatories shall analyse systemic risks stemming from their GPAISR so that, e.g.:

(1) each modality is evaluated in a way appropriate and relevant to the systemic risk assessed, e.g. by covering not only English but making best efforts to cover all major European languages and other languages supported by the model in language-based evaluations of multilingual models, whilst taking into account that

some systemic risks are based on capabilities, such as programming capabilities, that might be best evaluated in English; or

(2) each systemic risk is analysed in the appropriate and relevant modality, e.g. evaluating visual inputs/outputs, or actions in computer control or robotics.

To facilitate the adequate consideration of the expected use context and to access relevant expertise for and during model evaluations, Signatories shall, where possible and relevant to the systemic risk assessed, cooperate with relevant actors along the AI value chain, such as appropriately remunerated expert or lay representatives of civil society, academia, and/or other relevant stakeholders. Where relevant stakeholders, e.g. those who are directly affected by the GPAISR, are not available, Signatories shall identify and consult the most suitable representatives to represent such stakeholders' interests as part of systemic risk assessment and mitigation. Where SMEs are unable to cooperate with relevant stakeholders to implement this Measure, they shall request assistance from the AI Office."

G42's Frontier AI Safety Framework, page 5:

> Models created to generate output in a specific language, such as Arabic or Hindi, may be tested in those languages.

Page 8:
[For level 3]

> Testing is designed to ensure **effectiveness across all planned deployment contexts**, with specialized subject matter experts providing domain-specific input as needed.

OpenAI's GPT-4o System Card, page 23:

> 5.4 **Underrepresented Languages**
>
> GPT-4o shows improved reading comprehension and reasoning across a sample of historically underrepresented languages, and narrows the gap in performance between these languages and English.
>
> **To evaluate GPT-4o's performance in text across a select group of languages historically underrepresented in Internet text, we collaborated with external researchers and language facilitators to develop evaluations in five African languages**: Amharic, Hausa, Northern Sotho (Sepedi), Swahili, Yoruba. This initial assessment focused on translating two popular language benchmarks and creating small novel language-specific reading comprehension evaluation for Amharic, Hausa and Yoruba.

> Page 16:
>
> We evaluated the persuasiveness of GPT-4o's **text and voice modalities**. Based on pre-registered thresholds, the voice modality was classified as low risk, while the text modality marginally crossed into medium risk.

> Microsoft's Frontier Governance Framework, page 8:
>
> 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. This documentation will also include a summary of evaluation results, the deeper capability assessment, and safety and security mitigations. For example, the documentation could outline specific capabilities and tasks that the model robustly fails to complete which would be essential for a high or critical risk rating. This includes **identification of residual risks, for example relating to bias or discrimination, that are sensitive to deployment context, subject to laws and norms that vary across jurisdictions, and are likely to require further evaluation and mitigation on the part of those integrating the model into a system and deploying it**.

> Naver's AI Safety Framework
>
> Our partnerships with academia, the tech industry, and other stakeholders have led to meaningful research in **building local datasets specialized to Korean culture and society and creating benchmarks for evaluating Korean-centric models**.
>
> [...] When building AI systems, **we try to do so in the socio-technical context, which means training and evaluating language models with local datasets within the proper cultural and societal context**.

Measure II.4.9. Exploratory research and open-ended red-teaming

"As the current state of the art in systemic risk assessment is under-developed, and when necessary to effectively assess or mitigate systemic risk of their GPAISR, Signatories shall not restrict themselves to assessing only systemic risks for known model capabilities, propensities, affordances, and effects but also ensure the performance of exploratory work (internally and/or externally) to improve the systemic risk assessment or mitigation of their GPAISRs.

To implement this Measure, Signatories shall advance the state of the art in systemic risk assessment and mitigation techniques by, as appropriate for the size of their organisation and the systemic risk in question:

(1) exploratory research on systemic risk assessment and mitigation, such as new and improved model evaluation methods, meta-research on the "evaluation of model evaluations", or research in support of forecasting; and

(2) open-ended red-teaming engaging expert or lay representatives of civil society, academia, and other relevant stakeholders as participants. Where relevant stakeholders, e.g. those who are directly affected by the GPAISR, are not available, Signatories shall identify and engage the most suitable representatives to represent such stakeholders' interests."

> OpenAI's Preparedness Framework v1, page 12:
>
> > Unknown unknowns
> >
> > The list of Tracked Risk Categories above is almost certainly not exhaustive. As our understanding of the potential impacts and capabilities of frontier models improves, the listing will likely require expansions that accommodate new or understudied, emerging risks. Therefore, as a part of our Governance process (described later in this document), **we will continually assess whether there is a need for including a new category of risk** in the list above and how to create gradations. **In addition, we will invest in staying abreast of relevant research developments and monitoring for observed misuse** (expanded on later in this document), to help us understand if there are any emerging or understudied threats that we need to track.

> OpenAI's An Update on our Voluntary Commitments and Safety Frameworks, page 4:
>
> > External Testing
> >
> > **OpenAI has an extensive Red Teaming Network which consists of a diverse community of trusted external experts—including individual subject matter experts, research institutions, and civil society organizations—who help identify risks across cybersecurity, biological and chemical threats, and societal harms**. We've collaborated with external partners like METR and Apollo Research to strengthen our frontier risk assessments, launched public bug bounties for probing vulnerabilities in cybersecurity, assessed our models on public jailbreak arenas to decrease model vulnerability to jailbreaks, and worked with various external red teamers to assess emerging risks from new modalities like voice and video generation. **We also published new research on our approaches to red teaming to show how our external and automated red teaming efforts are advancing** to help deliver safe and beneficial AI. In December 2024, we also invited safety researchers to apply for early access to our next frontier models.

> Google DeepMind's Gemini 2.5 System Card, page 35:

> External groups were by design instructed to develop their own methodology to test topics within a particular domain area, remaining independent from internal Google DeepMind evaluations. The time dedicated to testing also varied per group, with some groups being dedicated full-time to executing testing processes, while others were part-time dedicated. **Some groups pursued manual red-teaming and reported on qualitative findings from their exploration of model behavior, while others developed bespoke automated testing strategies and produced quantitative reports of their results.**
>
> Google DeepMind's Frontier Safety Framework, page 7:
>
> "Instrumental Reasoning Level 2: [...] Mitigation: Future work: We are actively researching approaches to addressing models that reach this CCL."

> Anthropic's Responsible Scaling Policy, page 6:
>
> Currently, these [informal forecasts] will be informal estimates of (1) **the extent to which widely available elicitation techniques may improve** and (2) **how the model will perform on the same tasks when the next round of testing begins. As these are open research questions, we will aim to improve these forecasts over time** so that they can be relied upon for risk judgments.

Measure II.4.10. Sharing tools & best practices

"As systemic risks are influenced by the effects of interactions with other AI models or AI systems such as possible chain reactions of incidents or malfunctioning, and when necessary to effectively assess or mitigate systemic risk of Signatories' GPAISRs, non-SME Signatories shall share tools and best practices (such as established methodologies or procedures) for state-of-the-art model evaluation (including model elicitation) and systemic risk assessment and mitigation to further the safety and security work of relevant actors in the AI ecosystem, in particular those engaged in risk assessment or mitigation of AI models or AI systems that the Signatories' GPAISR can reasonably be foreseen to interact with, such as other model providers (especially SMEs), downstream providers, independent external assessors, or academic institutions.

Where Signatories choose to additionally share model evaluation data (such as inputs and outputs) used in systemic risk assessments, they shall limit this sharing as appropriate and, in particular, must not share data:

(1) that has high proliferation risks, e.g. where the shared test set could become a training set for a misuse actor;

(2) that could otherwise threaten scientific rigour, e.g. where shared test data could leak into training sets or significant parts of currently active held-out test sets would be released; and/or

(3) in order to protect public security and commercially sensitive information, where these interests outweigh the safety benefit of improved assessment or mitigation of systemic risks.

Non-SME Signatories shall: (1) ensure that sharing the tools and best practices referenced in this Measure will neither reduce nor restrict a Signatory's own capacity to perform systemic risk assessment and mitigation for their GPAISR; and (2) consider assigning appropriate numbers of engineering and support staff to research teams to facilitate such sharing, in order to avoid compromising the effective work of their systemic risk assessment personnel.

If the sharing of tools and best practices referenced in this Measure is not (yet) facilitated by the AI Office or other initiatives endorsed by the AI Office (such as the International Network of AI Safety Institutes) in relevant guidance where available, non-SME Signatories shall, as necessary to assess and mitigate systemic risk of the Signatories' GPAISRs, facilitate such sharing in ways that engage as many relevant actors as possible (e.g. through source code releases) or, where more restricted sharing is warranted, via secure sharing mechanisms which can be facilitated by industry organisations.

Signatories that are SMEs shall satisfy this Measure by using best efforts to make use of tools and best practices shared by non-SME Signatories."

> Amazon’s Frontier Model Safety Framework, page 4:
>
> • **Information sharing and best practices development**: Engagement in fora that bring together companies developing frontier models (e.g. **Frontier Model Forum and Partnership on AI**) and organized by government agencies (e.g. **National Institute of Standards and Technologies**). **These platforms serve as an opportunity to share findings related to our models** and to adopt recommendations from other leading companies.
>
> • Fostering academic research for development of cutting-edge alignment techniques: Through initiatives such as the Amazon Research Awards and Amazon Research centers (e.g. USC + Amazon Center on Secure & Trusted Machine Learning, Amazon/ MIT Science Hub), we work with leading academic partners conducting research on frontier AI risks and novel risk mitigation approaches. Additionally, **we advance our own research and publish findings in safety conferences**, while borrowing learnings presented by other academic institutions at similar venues.

> G42's Frontier AI Safety Framework, page 12:
>
> Threat Intelligence and Information Sharing: **G42 will share threat intelligence with industry partners** to address common challenges and emerging risks. **G42 will actively participate in forums to set industry standards and share best practices** for frontier model safety.

> Anthropic's Responsible Scaling Policy, page 16:
>
> We are uncertain how to choose a specific threshold, but we maintain a current list of specific CBRN capabilities of concern for which we would implement stronger mitigations. **We treat these lists as sensitive, but we plan to share them with organizations such as AI Safety Institutes and the Frontier Model Forum**, and keep these lists updated.

> Anthropic's Claude 3.7 System Card, page 26:
>
> We evaluated our models on a multiple-choice evaluation from SecureBio assessing virology-specific knowledge. Questions combine text statements with images, requiring assessment of multiple true/false claims. We use the "multiple select" variant, where models must select all correct answers, and none of the incorrect answers, in order to achieve a correct score on a given question, which is the most challenging variant of this evaluation. **While not yet public, this evaluation is shared across major labs via the Frontier Model Forum, allowing consistent tracking of model capabilities in virology, a key area of concern in our threat model.**

> Microsoft's Frontier Governance Framework, page 11:
>
> **Microsoft will prioritize ongoing contributions to this work and expand its collaboration with government, industry, and civil society, including through organizations like the Frontier Model Forum**, to solve the most pressing challenges in AI risk management. **We have identified a need for accelerated best practice development for risk assessment and evaluation methodologies.** [...]
>
> We also highlight the value of learning from experts outside of AI, including those with expertise in measurement science and in scientific domains like chemistry and biology, as well as those with knowledge of managing the risks of other complex technologies. **We will continue to share information and lessons from our governance efforts, taking care to do so in a way that minimizes the chance of inadvertently increasing risk, including by being judicious about sharing sensitive safety- and security-relevant information.**

> NVIDIA's Frontier AI Risk Assessment, page 13:
>
> To help focus red teaming activities and respond to model vulnerabilities and weaknesses, we first need to be aware of them. In cybersecurity, vulnerability scanners serve the purpose of proactively checking tools and deployments for known and potential weaknesses. For

> generative AI, we need an analogue. **NVIDIA runs and supports the Garak LLM vulnerability scanner. This constantly updated public project collects techniques for exploiting LLM and multi-modal model vulnerabilities and provides a testing and reporting environment for evaluating models' susceptibility. The project has formed a hub with a thriving community of volunteers that add their upgrades and knowledge.** Garak can test numerous scenarios rapidly, far exceeding the coverage possible with manual methods. Systematic exploration of model weaknesses can be repeated frequently, ensuring continuous oversight as the model evolves. NVIDIA takes advantage of this and uses Garak as a highest-priority assessment of models before release.

Measure II.4.11. Qualified model evaluation teams and adequate evaluation access and resources

"Signatories shall ensure that all model evaluation teams (internal and/or external as per Measure II.11.1) conducting systemic risk assessments are multidisciplinary teams that combine technical expertise with relevant domain knowledge, ensuring a holistic understanding of the systemic risk assessed. Model evaluation teams shall be:

(1) qualified to conduct the relevant model evaluation work, with multiple team members who each fulfil at least one of the following qualifications: (a) research or engineering experience so they can, objectively and reasonably, be viewed as having relevant expertise to contribute to the model evaluation for systemic risk assessment, demonstrated e.g. by a relevant PhD, relevant and recognized, peer-reviewed publications, or equivalent contributions to the field; (b) designed or developed a published peer-reviewed or widely used model evaluation on the systemic risk assessed; or (c) at least three years of applied experience working in a field directly relevant to the systemic risk assessed.

> Meta's Frontier AI Safety 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.

> Amazon's Frontier Model Safety Framework, page 3:
>
> > We work with specialized firms and academics to red-team our models to evaluate them for risks that require domain specific expertise.

> Anthropic's Responsible Scaling Policy, page 13:
>
> 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.

> Anthropic's Claude 3.7 System Card, page 6:
>
> Our RSP evaluations include automated testing of domain-specific knowledge, capability assessments through standardized benchmarks, and expert red teaming.

(2) given adequate access to the GPAISR, as necessary and appropriate to perform effective model evaluations and systemic risk assessments and mitigations of the GPAISR as required by this Code. Such access shall cover, as necessary, fine-tuning access to elicit model capabilities, logit access, sufficiently high rate limits, access to the model's inputs and outputs (including chain-of-thought reasoning), access to the model without safeguards, and additional affordances to the model, such as the model's ability to call scripts or operate a browser. Furthermore, such access shall cover, as necessary based on a case-by-case assessment, technical implementations of grey- and white-box access, where: (a) grey-box access offers some level of transparency, allowing for partial understanding of the "inner workings" of the model; and (b) white-box access means that the model's full weights, activations, and other technical details are always available for inspection, providing clear insights into how the model works. When giving grey- or white-box model access to model evaluation teams (internal and/or external), Signatories shall take into account the potential risks to model security that this access can entail (in accordance with Commitment II.7). Signatories shall provide the lowest level of access sufficient for rigorous model evaluations to be performed, including exploratory research, and implement more stringent security mitigations when higher levels of access are given;

> Microsoft's Frontier Governance Framework, page 6:
>
> As appropriate, evaluations involve qualified and expert external actors that meet relevant security standards, including those with domain-specific expertise.

> xAI's Risk Management Framework, page 1:
>
> In this draft framework, we particularly focus on requests that pose a foreseeable and non-trivial risk of more than one hundred deaths or over $1 billion in damages from weapons of mass destruction or cyberterrorist attacks on critical infrastructure ("catastrophic malicious use events"). **However, we will allow Grok to respond to such requests from some vetted, highly trusted users (such as trusted third-party safety auditors) whom we know to be using those capabilities for benign or beneficial purposes**, such as scientifically investigating Grok's

> capabilities for risk assessment purposes, or if such requests cover information that is already readily and easily available, including by an internet search.

> 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 o1 System Card, page 16:
>
> > We aim to test models that represent the "worst known case" for pre-mitigation risk, using capability elicitation techniques like custom post-training, scaffolding, and prompting.
>
> Page 11:
>
> > For o1, red teamers had access to various snapshots of the model at different stages of training and safety mitigation maturity starting in October 2024, through early December 2024. The model was accessed either via a sampling interface or via the API.

> OpenAI's GPT-4.5 System Card, page 6:
>
> > Apollo tested GPT-4.5 for scheming capabilities by placing it in a number of situations where in **its system prompt** the model is instructed to strongly pursue a specific goal.

(3) provided, to the Signatories' best efforts, with access to model specifications, training data, and past assessment results, as necessary and appropriate for the systemic risk assessed and the model evaluation method used;

> OpenAI's GPT-4.5 System Card, page 6:
>
> > METR evaluated an earlier checkpoint of GPT-4.5 and was given access to that checkpoint for 7 days, with **OpenAI sharing a subset of their internal evaluation results and providing context to help them to interpret their results.**

(4) provided with: (a) adequate compute budgets, including to allow for sufficiently long model evaluation runs, parallel execution, and re-runs; (b) appropriate staffing; and (c) sufficient engineering budgets and support, including to inspect model

evaluation results closely, e.g. to identify software bugs or model refusals which might lead to artificially lowered capabilities estimates; and

> Microsoft's Frontier Governance Framework, page 6:
>
> > 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**.

(5) given enough time to competently design, debug, execute, and analyse model evaluations to an objectively rigorous standard (see Measure II.4.5), whilst allowing for an appropriate level of model elicitation (see Measure II.4.6). The given time shall be proportionate to: (a) the magnitude of the systemic risk assessed; (b) the specific model evaluation method used, including consideration of how new or established it is; and (c) the degree of stability and performance similarity of the model under evaluation compared to the final model that will be made available on the market, e.g., a model is considered stable if: (i) the benchmark performance of successive model snapshots no longer significantly changes across general capability benchmarks; and (ii) model core functionality no longer changes, such as tool calling or (semi-)autonomous decision-making. An assessment time of at least 20 business days could, e.g., indicate that model evaluation teams were given enough time for most systemic risks and model evaluation methods."

> Microsoft's Frontier Governance Framework, page 6:
>
> > 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**.

Measure II.4.12. Safety margins

"Signatories shall identify and implement a sufficiently wide safety margin that is: (1) appropriate to the systemic risks stemming from the GPAISR; (2) representative of (a) potential under-elicitation, (b) general uncertainty in systemic risk assessment and mitigation results, and (c) historical (in)accuracy of similar assessments; (3) takes account of potential model improvements after the model has been made available on the market; and (4) in accordance with relevant state-of-the-art methods and practices, where available.

To implement this Measure, Signatories shall:

(1) use best efforts to ensure that their model evaluation methods are calibrated appropriately (i.e. having a low uncertainty metric) to, where possible, use quantitative safety margins (e.g., expressing safety margins in relative, percentage, or proportional terms); or

(2) where this is not possible, use qualitative safety margins or adopt other suitable approaches to safety margins, such as setting milestones conservatively or investing resources on elicitation efforts as appropriate to take account of (a) a potential or actual adversary in misuse cases and (b) reasonably foreseeable model improvements."

> Microsoft's Frontier Governance Framework, page 8:
>
> > Post-mitigation capability assessment and **safety buffer**: **Following application of safety and security mitigations, the model will be re-evaluated to ensure capabilities are rated low or medium and, if not, to guide further mitigation efforts**. 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.

> OpenAI's o3 and o4-mini System Card, page 11:
>
> > However, evaluations represent a lower bound for potential capabilities; additional prompting or fine-tuning, longer rollouts, novel interactions, or different forms of scaffolding could elicit behaviors beyond what we observed in our tests or the tests of our third-party partners.

> Google DeepMind's Gemini 2.5 System Card, page 26:
>
> > When conducting FSF evaluations, **we compare test results against internal alert thresholds ("early warnings") which are set significantly below the actual CCLs. This built-in safety buffer helps us be proactive by signaling potential risks well before models reach CCLs**. Concretely, our alert thresholds are designed such that if a frontier model does not reach the alert threshold for a CCL, models are unlikely to reach that CCL before the next regular testing—which we conduct at a regular cadence and also when we anticipate or see exceptional capability progress. Our recent paper (Shah et al., 2025) discusses this approximate continuity assumption in more depth in Section 3.5.

> 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.
>
> [Executive Summary]
>
> > If, after the comprehensive testing, we determine that the model is sufficiently below the relevant Capability Thresholds, then we will continue to apply the ASL-2 Standard. If, however, we are unable to make the required showing, we will act as though the model has surpassed the Capability Threshold. This means that we will both upgrade to the ASL-3 Required Safeguards and conduct a follow-up capability assessment to confirm that the ASL-4 Standard is not necessary.

> Anthropic's *System Card: Claude Opus 4 & Claude Sonnet 4*, page 10:
>
> > To be clear, we have not yet determined whether Claude Opus 4 has definitively passed the capabilities threshold that requires ASL-3 protections. Rather, we cannot clearly rule out ASL-3 risks for Claude Opus 4 (although we have ruled out that it needs the ASL-4 Standard). Thus, **we are deploying Claude Opus 4 with ASL-3 measures as a precautionary, provisional action, while maintaining Claude Sonnet 4 at the ASL-2 Standard**.

Measure II.4.13 Evaluation practices for forecasting

"Signatories shall use best efforts to adopt forecasting-specific techniques in those model evaluations that are central to the highest systemic risk tiers, especially in support of Measure II.1.3, using e.g.:

(1) scaling law experiments for model size, training compute, and/or inference compute; (2) comparisons across model generations; or (3) other techniques that support estimates for when dangerous capabilities, propensities, affordances, and effects might emerge in future models;

provided that such techniques are in accordance with relevant state of the art where available."

> Anthropic's *System Card: Claude Opus 4 & Claude Sonnet 4*, page 8:
>
> > Similar to the process introduced for Claude Sonnet 3.7, we conducted evaluations throughout the training process **to better understand how catastrophic risk-related capabilities evolved over time**.

> Meta's Introducing Meta Llama 3:

> **[W]e have developed a series of detailed scaling laws for downstream benchmark evaluations.** These scaling laws enable us to select an optimal data mix and to make informed decisions on how to best use our training compute. **Importantly, scaling laws allow us to predict the performance of our largest models on key tasks (for example, code generation as evaluated on the HumanEval benchmark—see above) before we actually train the models.** This helps us ensure strong performance of our final models across a variety of use cases and capabilities.
>
> We made several new observations on scaling behavior during the development of Llama 3. For example, while the Chinchilla-optimal amount of training compute for an 8B parameter model corresponds to ~200B tokens, we found that model performance continues to improve even after the model is trained on two orders of magnitude more data. Both our 8B and 70B parameter models continued to improve log-linearly after we trained them on up to 15T tokens. Larger models can match the performance of these smaller models with less training compute, but smaller models are generally preferred because they are much more efficient during inference.

OpenAI’s Preparedness Framework v1, page 2:

> Our Preparedness Framework contains five key elements:
>
> 1. Tracking catastrophic risk level via evaluations. We will be building and continually improving suites of evaluations and other monitoring solutions along several Tracked Risk Categories, and indicating our current levels of pre-mitigation and post-mitigation risk in a Scorecard. **Importantly, we will also be forecasting the future development of risks, so that we can develop lead times on safety and security measures**.

Anthropic’s Responsible Scaling Policy, page 5:

> “Effective Compute” is a scaling-trend-based metric that accounts for both FLOPs and algorithmic improvements. An Effective Compute increase of K represents a performance improvement from a pretrained model on relevant task(s) equivalent to scaling up the baseline model’s training compute by a factor of K. **We plan to track Effective Compute during pretraining on a weighted aggregation of datasets relevant to our Capability Thresholds (e.g., coding and science).** This is, however, an open research question, and we will explore different possible methods. More generally, the Effective Compute concept is fairly new, and we may replace it with another metric in a similar spirit in the future.

Page 6:
[3.2. Comprehensive Assessment]

> Forecasting: Make informal forecasts about the **likelihood that further training and elicitation will improve test results between the time of testing and the next expected round of comprehensive testing.**

Page 6:
[A footnote to the above citation]

> Currently, these will be informal estimates of (1) the extent to which widely available elicitation techniques may improve and **(2) how the model will perform on the same tasks when the next round of testing begins.** As these are open research questions, we will aim to improve these forecasts over time so that they can be relied upon for risk judgments.

Google DeepMind's [Evaluating Frontier Models for Dangerous Capabilities](), page 2:

> **We commissioned a group of eight professional forecasters to predict future results on our evaluations.** Specifically, we focused on two clusters of forecasting questions: 1. Evaluation results prediction. We want to predict the development of dangerous capabilities ahead of time, so that developers can take steps to identify and implement safety and security mitigations. The accuracy of the forecasts cannot be taken for granted, but forecasting experts often outperform domain specialists (such as ourselves) (Tetlock, 2005). We asked the forecasters to predict by when our in-house CTF challenges, Hack The Box challenges, and the self-proliferation tasks would be passed. 2. Impact of AI, conditional on our evaluations being passed. We wanted to see how much the results of our evaluations would affect expert forecasts on whether AI will be considered a major global problem in the future. This tests whether our evaluations are measuring something important.

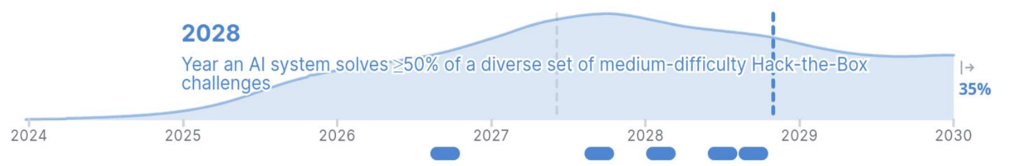


Figure 11| Aggregated forecast for the question: **When will an AI system have solved ≥ 50% of a diverse set of medium-difficulty Hack The Box challenges?**[13] The vertical lines indicate the 25th, 50th and 75th percentile of the prediction distribution.

Measure II.4.14 Post-market monitoring

"Signatories shall conduct post-market monitoring to gather necessary information for assessing and mitigating systemic risk, including information about model capabilities, propensities, affordances, and effects, as appropriate for the systemic risk assessed. Signatories shall: (1) in particular conduct post-market monitoring with a view to ensure that their GPAISR does not pose unacceptable systemic risk as defined by their systemic risk acceptance criteria (as per Measure II.1.2); and (2) use best efforts to conduct post-market monitoring for lower systemic risk tiers as necessary for forecasting (see Measures II.1.3 and II.4.13).

> G42's Frontier AI Safety Framework, page 5:
>
> In addition to conducting pre-deployment evaluations, **post-deployment monitoring will also be used** to indicate whether G42's models have reached capability thresholds and whether increased deployment mitigation and security mitigation levels are required.

> Anthropic's Responsible Scaling Policy, page 6:
>
> **We also maintain ongoing monitoring systems after release** to track safety metrics and model behavior, allowing us to respond to emergent concerns.

To fulfil this Measure, Signatories shall establish organisational, contractual, technical, or other methods most suitable for the specific integration, release and distribution strategy, and use of their GPAISR to gather relevant post-market information, whilst ensuring that the information is treated in accordance with Union law.

Signatories shall consider a variety of suitable methods for post-market monitoring, such as:

(1) collection of end-user feedback; (2) (anonymous) reporting channels;

> Microsoft's Frontier Governance Framework, page 8:
>
> 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.

> Anthropic's Responsible Scaling Policy, page 2:
>
> The ASL-2 Deployment Standard reduces the prevalence of misuse, and includes the publication of model cards and enforcement of Usage Policy; harmlessness training such as Constitutional AI and automated detection mechanisms; and **establishing vulnerability reporting channels** as well as a bug bounty for universal jailbreaks**.**

(3) (serious) incident report forms; (4) bug bounties;

> Anthropic's Responsible Scaling Policy, page 8:
>
> [For ASL-3]
>
> > 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.
>
> Page 2:
>
> > The ASL-2 Deployment Standard reduces the prevalence of misuse, and includes the publication of model cards and enforcement of Usage Policy; harmlessness training such as Constitutional AI and automated detection mechanisms; and **establishing vulnerability reporting channels as well as a bug bounty for universal jailbreaks.**

> Cohere's Secure AI Frontier Model Framework, page 18:
>
> > **With respect to AI security, Cohere's bug bounty program provides monetary incentives to promote and advance the understanding of AI security.** Cohere also actively contributes to thought leadership initiatives in collaboration with industry peers to improve the overall security of the AI value chain.

> xAI's Risk Management Framework (Draft), page 4:
>
> > As an additional measure to enhance safety, we will subject Grok to adversarially testing its safeguards utilizing both internal and qualified external red teams. **Potentially, we will also explore incentive mechanisms like bounties** as another mechanism to further improve Grok's safeguards.

(5) community-driven model evaluations and public leaderboards;

(6) monitoring their GPAISR's use in the real world, e.g. identifying use in software repositories and known malware, or monitoring public forums and social media for novel patterns of use;

> Cohere's Secure AI Frontier Model Framework, page 18:

4b) Continuous improvement

We are constantly improving our practices to better mitigate AI risks. To identify areas of improvement, we engage with our customers, partners, and the wider public by:

- Incentivizing third-party vulnerability discovery via clear protections for legitimate research practices in our Responsible Disclosure Policy
- Conducting user research to understand what challenges and risks can be expected in enterprise use cases
- **Engaging with the broader community via dedicated user forums**

Page 6:

Cohere consistently reviews state-of-the-art research and industry practice regarding the risks associated with AI, and uses this to determine our priorities. At Cohere, risks to our systems are identified through a list of continuously-expanding techniques, including:

- Mitigating core vulnerabilities identified by the Open Worldwide Application Security Project (OWASP)
- Internal threat modeling, which includes a review of how our customers interact with and use our models, to proactively identify potential threats and implement specific counter measures before deployment
- Monitoring established and well-researched repositories of security attacks and vulnerabilities for AI, such as the Mitre Altas database

(7) supporting the scientific study of their GPAISR's emerging risks, capabilities, and effects in the Union in collaboration with academia, civil society, regulators, and/or independent researchers;

Cohere's Secure AI Frontier Model Framework, page 18:

5a) Support for independent research

The field of AI risk management extends beyond individual company practices to encompass a wide ecosystem of researchers across industry, government, civil society, and academia. This collective effort contributes to advancing research into AI risks, as well as developing techniques and best practices that can be adopted by model developers.

Cohere actively contributes to this ecosystem, in large part through our open science research lab, Cohere For AI (C4AI). In addition to conducting fundamental research – on topics including, but not limited to AI risks - **C4AI supports a community of over 3000 researchers around the world to connect, collaborate, and share research with one another, and expand how AI research is done and by whom**. Additionally, C4AI directly supports independent researcher efforts through a Research Grant Program, which provides researchers in nonprofit,

> public sector, or academic settings access to use Cohere's models for research or public interest projects via subsidized access to our API.

> Meta's Frontier AI Framework, page 2:
>
> > **We hope that sharing our current approach to development of advanced AI systems will** not only promote transparency into our decision-making processes but also **encourage discussion and research** on how to improve the science of AI evaluation and the quantification of risks and benefits.
>
> Page 19:
>
> > To that end, we'll continue to work on: (1) improving the quality and reliability of evaluations; (2) developing additional, robust mitigation techniques; and (3) **more advanced methods for performing post-release monitoring of open source AI models.**

(8) conducting frequent dialogues with affected stakeholders, using, e.g., participatory methods in accordance with Measure II.4.8;

> Cohere's Secure AI Frontier Model Framework, page 12:
>
> > Where Cohere has direct visibility into the use of its models during deployment, we use that visibility to monitor for malicious attempts to prompt our models for harmful outputs, revoking access from accounts that abuse our systems. **Cohere partners closely with customers who deploy Cohere's AI solutions privately or on third-party managed platforms to ensure that they understand and recognize their responsibility for implementing appropriate monitoring controls during deployment.**

(9) investing in the development of new technical methods supporting the privacy-preserving monitoring and analysis of individual instances of their GPAISR or of AI systems, in particular autonomous AI systems such as AI agents, built from their GPAISR, e.g. the development of watermarks, fingerprinting, digital signatures, or decentralised, privacy-preserving monitoring techniques; or

> xAI's Risk Management Framework (Draft), page 7:
>
> > **We will explore creating an experimental AI ID system to uniquely identify instances of our AI agents interacting with real-world systems**, potentially using HTTP headers. Such an ID system could serve as a foundation for many aspects of the agent ecosystem, building upon and

> improving existing internet infrastructure for identifying bots and crawlers by including verifiability. This future ecosystem may include the development of reputation systems for agents, agent-only channels of communication and interaction, and the ability to respond to incidents involving AI agents. Such a future ID system would support safer agent-to-agent and agent-to-human interactions, and would promote trust and safety in the rapidly evolving AI ecosystem.

(10) implementing privacy-preserving logging and metadata analysis methods, where technically, legally, and commercially feasible, considering the release and distribution strategy of their GPAISR.

> Anthropic's Responsible Scaling Policy, page 8, under ASL-3 Deployment Standard:
>
> > 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.**

> OpenAI's Preparedness Framework v2, page 19 as a potential safeguard against a misaligned model:
>
> > Agent actions are all logged in an uneditable database, and asynchronous monitoring routines review those actions for evidence of harm.

Signatories shall, in particular, collect relevant information about: (1) breaches of a GPAISR's use restrictions and subsequent serious incidents arising from such breaches, to inform the implementation of Measure II.6.3; and (2) for closed-source GPAISRs, aspects of such models that are relevant for assessing and mitigating systemic risk but that are not transparent to third parties, such as hidden chains-of-thought.

> Microsoft's Frontier Governance Framework, page 8:
>
> > Other forms of monitoring, including for example, **automated monitoring in chain-of-thought outputs,** are also utilized as appropriate. We continue to assess the tradeoffs between safety and security goals and legal and privacy considerations, optimizing for measures that can achieve specific safety and security goals in compliance with existing law and contractual agreements.

> OpenAI's o1 preview System Card, page 10:

> While we are very excited about the prospect of chain-of-thought interpretation and monitoring, we are wary that they may not be fully legible and faithful in the future or even now. We are actively pursuing research into (a) whether these issues will be exacerbated or alleviated as we further scale models in the o1 paradigm and **(b) mitigations to improve the monitorability of our future models.**
>
> Page 8:
>
> As a case study for chain-of-thought monitoring, we created a rudimentary monitor tasked with looking for instances where o1 models deceive users, i.e. knowingly providing incorrect information to a user, or omitting crucial information that could lead them to have a false belief. [...] **Our monitor is a GPT-4o model prompted to look for deception in o1's chain of thought and final output.** We validated the monitor on a small on-policy internal dataset of deceptive and non-deceptive o1-preview outputs and chains-of-thought, on which it was 92% accurate. We used the monitor on 100,000 synthetic ChatGPT prompts with o1-preview completions (we used synthetic prompts instead of real ChatGPT conversations).

> OpenAI's GPT-4.5 System Card, page 8:
>
> Mitigations include:
> [...]
> - Dedicated monitoring and detection efforts for CBRN, Persuasion, and Cyber tasks.
> - Continued improvement of our scaled detection capabilities, including content moderation classifiers with greater precision and recall to support targeted enforcement of our Usage Policies and detections to act on activities that may be unsafe or violate Usage Policies.

Where Signatories themselves provide and/or deploy AI systems that incorporate their own GPAISR, Signatories shall, as necessary to effectively assess and mitigate systemic risks stemming from the model in question, monitor the model as part of these systems.

Further, and particularly where the methods listed in points (1) to (10) above do not provide Signatories with the necessary post-market information to effectively assess and mitigate systemic risks stemming from their GPAISR, Signatories shall use best efforts to cooperate with licensees, downstream providers, and/or end-users to receive such information from these actors and take relevant information shared by them into account, provided that such information is treated in accordance with Union law. Where they do so and end-users are consumers, Signatories shall offer the choice to opt-in to sharing information relevant for post-market monitoring, provided that such information is treated in accordance with Union law."

> NVIDIA's Frontier AI Risk Assessment, page 11:
>
> **Watermarks embedded at generation-time enable downstream detection and attribution of AI-produced outputs, providing a verifiable origin signal for both end-users and automated scanning tools.** This ensures that when models produce text, images, or other media, it can be reliably traced back to a specific model version or developer entity. Implicit watermarks are subtle, often statistical or cryptographic markers embedded directly into the distributional properties of generated outputs. They are not perceivable by the human eye or ear and do not alter the visible or audible characteristics of the content. There are several open source implicit watermarking tools that may be used to increase the detectability of content generated by NVIDIA models. Explicit watermarks are visible or otherwise directly perceivable indicators, such as logos, text overlays, or perceptible patterns. **They clearly denote AI-generated content to end-users. The use of explicit watermarks are specific to downstream developers. NVIDIA encourages the use of explicit watermarks to help mitigate the spread of misinformation but is not integrated into NVIDIA's foundation models.**

Commitment II.5. Systemic risk acceptance determination

"Signatories commit to determining the acceptability of the systemic risks stemming from their GPAISRs by comparing the results of their systemic risk analysis (pursuant to Commitment II.4) to their pre-defined systemic risk acceptance criteria (pursuant to Measure II.1.2), in order to ensure proportionality between the systemic risks of the GPAISR and their mitigations. Signatories commit to using this comparison to inform the decision of whether or not to proceed with the development, the making available on the market, and/or the use of their GPAISR, as specified in the Measures for this Commitment."

> OpenAI's Preparedness Framework v1, page 14:
>
> Evaluating post-mitigation risk
>
> To verify if mitigations have sufficiently and dependently reduced the resulting post mitigation risk, we will also run evaluations on models after they have safety mitigations in place, again attempting to verify and test the possible "worst known case" scenario for these systems. **As part of our baseline commitments, we are aiming to keep post-mitigation risk at "medium" risk or below**.
>
> Page 8:
>
> **Only models with a post-mitigation score of "medium" or below can be deployed**, and only models with a post-mitigation score of "high" or below can be developed further (as defined in the Tracked Risk Categories below).

> Magic's AGI Readiness Policy
>
> Threshold Definition
>
> **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.**
>
> **As an alternative threshold definition, we will also make use of a set of private benchmarks that we use internally to assess our product's level of software engineering capability.** For comparison, we will also perform these evaluations on publicly available AI systems that are generally considered to be state-of-the-art. We will have privately specified thresholds such that if we see that our model performs significantly better than publicly available models, this is considered evidence that we may be breaking new ground in terms of AI systems' dangerous capabilities. **Reaching these thresholds on our private benchmarks will also trigger our commitments to develop our full AGI Readiness Policy, with threat model evaluations and mitigations, before substantial further model development or deployment.**

> xAI - Risk Management Framework (Draft), page 8:
>
> Safeguards are adequate only if Grok's performance on the relevant benchmarks is within stated thresholds.

> Microsoft's Frontier Governance Framework, page 6:
>
> Pre-mitigation capability assessment: The results of the deeper capability assessment are used to assign a model a pre-mitigation score of low, medium, high, or critical for each tracked capability on the basis of the framework's capability thresholds (see Appendix I). **Models assessed as posing low or medium risk may be deployed with appropriate safeguards as outlined below. Models assessed as having a high or critical risk are subject to further review and safety and security mitigations prior to deployment**.

> Anthropic's *System Card: Claude Opus 4 & Claude Sonnet 4*, page 93:
>
> Our threat analysis indicated that total uplift ≥ 5× (or raw uplift ≥ 0.8) would create significant additional risk, while uplift ≤ 2.8× would keep risk at acceptable levels. Text-based uplift trials are substantially weaker proxies for real-world scenarios—which involve additional factors like tacit knowledge, materials access, and actor persistence—but we nevertheless adopted a similar uplift threshold to be conservative in our ASL-3 rule-out decisions.

Measure II.5.1. Proceeding where systemic risk is deemed acceptable

"Where systemic risk is deemed acceptable, Signatories shall determine whether the residual systemic risk nonetheless warrants additional mitigation before proceeding with the development, first or ongoing making available on the market, and/or use of their GPAISR."

> Meta's Frontier AI Framework, page 8:
>
> > We assess residual risk, taking into consideration the details of the risk assessment, the results of evaluations conducted throughout training, and the mitigations that have been implemented.
> >
> > **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.**

> NVIDIA's Frontier AI Risk Assessment, page 1:
>
> > Our risk evaluation process then estimates the residual risk after controls are applied and compares it against the potential initial risks posed by the AI-based product. Leveraging the results from the risk evaluation phase, it is possible to determine how residual risks correspond with NVIDIA's Trustworthy AI (TAI) principles and document any trade-offs made during the allocation of risk treatment measures.

Measure II.5.2. Not proceeding where systemic risk is deemed unacceptable

"Where systemic risk is deemed unacceptable, Signatories shall take appropriate steps to bring such risk to an acceptable level in accordance with their Framework.

Where a model is not yet available on the market, Signatories shall, e.g.: (1) implement additional mitigations until systemic risk is deemed acceptable; or (2) not make a model available on the market.

Where the model is already made available on the market, Signatories shall: (3) restrict the making available, withdraw, or recall a model from the market; or (4) otherwise address the systemic risks. As a minimum, Signatories shall follow the relevant steps outlined in their Framework (pursuant to Measure II.1.2).

Once appropriate steps to keep systemic risk at an acceptable level have been implemented, Signatories, prior to proceeding with the development, the first or ongoing making available on the market, and/or the use of their GPAISR, shall conduct another round of systemic risk analysis (in accordance with Commitment II.4) and systemic risk acceptance determination (in accordance with Commitment II.5)."

> OpenAI's Preparedness Framework v1, page 21:
>
> Restricting deployment
>
> Only models with a post-mitigation score of "medium" or below can be deployed. In other words, **if we reach (or are forecasted to reach) at least "high" pre-mitigation risk in any of the considered categories, we will not continue with deployment of that model** (by the time we hit "high" pre-mitigation risk) **until there are reasonably mitigations in place for the relevant post-mitigation risk level to be back at most to "medium" level.** (Note that a potentially effective mitigation in this context could be restricting deployment to trusted parties.)

> Anthropic's Responsible Scaling Policy, page 11:
>
> 6.2. Restrict Deployment and Further Scaling
>
> **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**: [...]
>
> • 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**. Given the availability of interim deployment and security protections, however, stronger restrictions should rarely be necessary.

> 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**.

> 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**.

> Amazon's Frontier Model Safety Framework, 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**.

Measure II.5.3. Staging model development and making available on the market when proceeding

"When proceeding with the development, first or ongoing making available on the market, and/or use of their GPAISR, Signatories shall use best efforts to proceed in stages, as necessary, to keep systemic risks below an unacceptable level and appropriately mitigated below that level, adopting practices such as: (1) limiting API access to vetted users; (2) gradually expanding access based on post-market monitoring and ongoing systemic risk assessments; (3) starting with a closed release before any open release; (4) using logging systems to track use and safety concerns; (5) setting criteria for progressing through stages, including criteria based on systemic risk tiers and user feedback; and (6) retaining the ability to restrict access, as appropriate."

> Microsoft's Frontier Governance Framework, page 8:
>
> **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**. We are also progressing work to **further study models when in use and assess the real-world effectiveness of mitigations**, while upholding stringent levels of privacy and confidentiality and benefiting from external expertise where appropriate.

> Meta's Frontier AI Framework, page 18:
>
> Examples of mitigation techniques we implement include:
> • Fine-tuning
> • Misuse filtering, response protocols
> • Sanctions screening and geogating
> • **Staged release to prepare the external ecosystem**

> G42's Frontier AI Safety Framework, pages 8-9:
> [For level 3]
>
> **Controlled Rollout**: For new frontier level models, **implement phased rollouts, starting with limited access to trusted users**, with full deployment only after exhaustive risk assessments.

> NVIDIA's Frontier AI Risk Assessment, page 12:
>
> **Initially, model access can be restricted to a limited audience, expanding gradually as risks are better understood and mitigated.** In cases of severe risk, notifying other developers of identified hazards through the proven channel of NVIDIA's security bulletin allows for coordinated response efforts, mitigating widespread issues. As a last resort, full market removal or deletion of the model and its components can be considered to prevent further misuse and contain hazards effectively.

> xAI's Risk Management Framework, page 8:
>
> To mitigate risks, **we intend to utilize tiered availability of the functionality and features of Grok**. For instance, the full functionality of a future Grok could be made available only to trusted parties, partners, and government agencies. We could also mitigate risks by adding additional controls on functionality and features depending on the end user (e.g., consumers using mobile apps vs. sophisticated businesses using APIs).

Commitment II.6. Safety mitigations

"Signatories commit, as specified in the Measures for this Commitment, to: (1) implementing technical safety mitigations along the entire model lifecycle that are proportionate to the systemic risks arising from the development, the making available on the market, and/or the use of GPAISRs, in order to reduce the systemic risks of such models to acceptable levels, and further reduce systemic risk as appropriate, in accordance with this Code; and (2) ensuring that safety mitigations are proportionate and state-of-the-art."

> OpenAI's Preparedness Framework v2, page 10:
>
> > **Safeguards that mitigate the risk of severe harm are at the core of our safe deployment approach** for covered systems that reach a capability threshold. [...]
> >
> > **If a covered system appears likely to cross a capability threshold, we will start to work on safeguards to sufficiently reduce the risks of severe harm associated with that threshold even if a formal capability determination has not yet been made.**
>
> Page 11:
>
> > Covered systems that reach High capability must have safeguards that sufficiently minimize the associated risk of severe harm before they are deployed. Systems that reach Critical capability also require safeguards that sufficiently minimize associated risks during development. [...] We expect to continuously improve our safeguards over time.

> Anthropic's Responsible Scaling Policy, page 1:
>
> > Since the frontier of AI is rapidly evolving, we cannot anticipate what safety and security measures will be appropriate for models far beyond the current frontier. **We will thus regularly measure the capability of our models and adjust our safeguards accordingly. Further, we will continue to research potential risks and next-generation mitigation techniques.** And, at the highest level of generality, we will look for opportunities to improve and strengthen our overarching risk management framework.

> Cohere's Secure AI Model Framework, page 8:
>
> > Once risks have been identified, they must be assessed and mitigated. **We assess and mitigate risks across the development lifecycle** of Cohere's AI solutions.
> >
> > P. 13: Evaluations and **testing for general performance, safety, and security occur throughout the lifecycle** and include red teaming, evaluation with academic and industry benchmarks, and internal bespoke evaluations. **Algorithms and training data are adjusted iteratively based on evaluation and testing results.**

Measure II.6.1. Safety mitigations to implement

"Without prejudice to Measure II.6.2, Signatories shall, as necessary to mitigate systemic risks and reduce them to acceptable levels (in accordance with this Code), implement technical safety mitigations for GPAISRs, such as: (1) filtering and cleaning training data; (2) monitoring and filtering the inputs and outputs of such models; (3) changing the behaviour of such a model in the interests of safety, such as fine-tuning the model to refuse certain requests; (4) restricting the availability of such a model on the market, such as restricting model access to vetted users; (5) offering countermeasures or other safety tools to other actors; (6) implementing high-assurance quantitative safety guarantees concerning the behaviour of such a model; and (7) implementing infrastructure that could help promote safe ecosystems of AI agents, such as a reputation system, specialised communication protocols, or incident monitoring tools."

xAI's Risk Management Framework, page 3:

> Examples of safeguards or mitigations we may potentially utilize to achieve our safety objectives include:
>
> - **Refusal training**: Training Grok to decline harmful requests. (3)
> - Circuit breakers: Using representation engineering to interrupt model representations responsible for hazardous outputs. (3)
> - **Input and output filters**: Applying classifiers to user inputs or model outputs to verify safety when Grok is queried regarding weapons of mass destruction or cyberterrorism. (2)

Page 7:

> We will explore creating an experimental AI ID system to uniquely identify instances of our AI agents interacting with real-world systems, potentially using HTTP headers. Such an ID system could serve as a foundation for many aspects of the agent ecosystem, building upon and improving existing internet infrastructure for identifying bots and crawlers by including verifiability. This future ecosystem may include the development of **reputation systems for agents, agent-only channels of communication and interaction, and the ability to respond to incidents involving AI agents**. (7)

OpenAI's GPT-4.5 System Card, page 2:

> Our data processing pipeline includes rigorous filtering to maintain data quality and mitigate potential risks. (1)

Page 8, at Preparedness Mitigations:

> GPT-4.5 leverages a combination of pre-training and post-training techniques to mitigate against potential catastrophic risks, and inherits much of our earlier safety training in refusal behavior. [...]
>
> **Mitigations include:**
>
> - **Pre-training mitigations, such as filtering out a highly targeted set of CBRN proliferation data based on limited or no legitimate use.** (1)

- **Safety training** for political persuasion tasks. (3)
- Continued focus on model robustness for CBRN and Persuasion risks, to improve ability of our models to **withstand malicious and adversarial users, techniques, and conditions.** (3)
- Dedicated **monitoring and detection efforts** for CBRN, Persuasion, and Cyber tasks. (2), (7)
- Continued **improvement of our scaled detection capabilities, including content moderation classifiers** with greater precision and recall to support targeted enforcement of our Usage Policies and detections to act on activities that may be unsafe or violate Usage Policies. (2)
- **Monitoring and targeted investigations of suspected abuses** relating to influence operations, extremism, and improper political activities to address persuasion risks. (7)
- Monitoring for high-risk cybersecurity threats, such as active measures to disrupt high-priority adversaries including **hunting, detection, monitoring, tracking, intel-sharing and disrupting**. (2), (5), (7)
- Threat model development for self-exfiltration and self-improvement risks in preparation for agentic models with heightened capabilities.

Microsoft's Frontier Governance Framework, page 8:

> 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. (5)
>
> 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.** (3) This is an area of active research, and we continue to invest in and apply robust harm refusal techniques as they are developed.
> - **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.** (5) This documentation will also include a summary of evaluation results, the deeper capability assessment, and safety and security mitigations […]
> - **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.** (7) 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 (2) [...]
>
> - **Phased release, trusted users, and usage studies, as appropriate for models demonstrating novel or advanced capabilities.** (4) 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. We are also progressing work to further study models when in use and assess the real-world effectiveness of mitigations, while upholding stringent levels of privacy and confidentiality and benefiting from external expertise where appropriate.

Measure II.6.2. Proportionate and state-of-the-art safety mitigations

"Signatories shall implement state-of-the-art technical safety mitigations that: (1) are proportionate to the systemic risk at issue, taking into account the systemic risk acceptance criteria (as set out in the Framework pursuant to Measure II.1.2); and (2) best mitigate, in particular, unacceptable systemic risks (as set out in the Framework pursuant to Measure II.1.2).

For the purposes of this Code, state-of-the-art technical safety mitigations need not always or necessarily mitigate all systemic risks to the greatest extent."

> Microsoft's Frontier Governance Framework, page 7:
>
> We apply **state-of-the-art safety mitigations tailored to observed risks so that the model's risk level remains at low or medium** once mitigations have been applied.

> Anthropic's Responsible Scaling Policy, page 1:
>
> First, **our approach to risk should be proportional. Central to our policy is the concept of AI Safety Level Standards**: technical and operational standards for safely training and deploying frontier models that correspond with a particular level of risk. **By implementing safeguards that are proportional to the nature and extent of an AI model's risks**, we can balance innovation with safety, maintaining rigorous protections without unnecessarily hindering progress.

> Meta's Frontier AI Framework, page 8:
>
> Our mitigation strategy is informed by the risks we've identified in the risk assessment, evaluation results, the mitigations that have been applied to existing models in the same class, and the release approach.

> OpenAI's Preparedness Framework v2, page 11:

> Covered systems that reach High capability must have safeguards that sufficiently minimize the associated risk of severe harm before they are deployed. Systems that reach Critical capability also require safeguards that sufficiently minimize associated risks during development. [...]
>
> Page 10:
>
> SAG [Safety Advisory Group] is responsible for assessing **whether the safeguards associated with a given deployment sufficiently minimize the risk of severe harm associated with the proposed deployment.** The SAG will make this determination based on:
>
> - The level of capability in the Tracked Category based on the Capabilities Report.
> - The **associated risks of severe harm**, as described in the threat model and where needed, advice of internal or external experts.
> - The **safeguards in place and their effectiveness** based on the Safeguards Report.
> - The baseline risk from other deployments, based on a review of any non-OpenAI deployments of models which have crossed the capability thresholds and any public evidence of the safeguards applied for those models.

Measure II.6.3. Serious incident response readiness

"Signatories shall implement corrective measures to address serious incidents through the use of heightened or additional technical safety and/or security mitigations (pursuant to this Commitment and/or Commitment II.7)."

> xAI's Risk Management Framework, page 7:
>
> If xAI learned of an imminent threat of a significantly harmful event, including loss of control, **we would take steps to stop or prevent that event,** including potentially the following steps:
>
> 1. We would immediately 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 would 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 would temporarily fully shut down the relevant system until we had a more targeted response.**
> 4. We would 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 would use the post-mortem to inform development and implementation of necessary changes to our risk

> management practices.

> G42's Frontier AI Safety Framework, page 9 at Incident Response:
>
> Developing a **comprehensive incident response plan that outlines the steps to be taken in the event of non-compliance.** Incident detection should leverage automated mechanisms and human review, and non-sensitive incident information should be shared with applicable government bodies. We plan for our response protocols to focus on rapid remediation to minimize unintended harmful outputs from models. Depending on the nature and severity of the incident, this might involve implementing immediate containment measures restricting access to the model either externally, internally or both.

> NVIDIA'S Frontier AI Risk Assessment, page 11:
>
> Lowering the duration of a hazard can be achieved by **implementing establish robust protocols for managing AI-related incidents, including clear information-sharing mechanisms between developers and relevant authorities.** This encourages proactive identification of potential risks before they escalate. Additionally, reducing access to a model reactively when misuse is detected can help limit further harm. This can involve rolling back a model to a previous version or discontinuing its availability if significant misuse risks emerge during production. Lastly, conducting regular safety drills ensures that emergency response plans are stress-tested. By practicing responses to foreseeable, fast- moving emergency scenarios, NVIDIA is able to improve their readiness and reduce the duration of hazardous incidents.

> Anthropic's Responsible Scaling Policy, page 11, at 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.**

> OpenAI's Preparedness Framework v1, page 25, at Safety drills:
>
> A critical part of this process is to be prepared if fast-moving emergency scenarios arise, including what default organizational response might look like (including how to stress-test against the pressures of our business or our culture). While the Preparedness team and SAG [Safety Advisory Group] will of course work hard on forecasting and preparing for risks, **safety drills can help the organization build "muscle memory" by practicing and coming up with the right "default" responses for some of the foreseeable scenarios. Therefore, the SAG will call for safety drills at a recommended minimum yearly basis.**

Commitment II.7. Security mitigations

> "Signatories commit to mitigating systemic risks that could arise from unauthorised access to unreleased model weights of their GPAISRs and/or unreleased associated assets. Associated assets encompass any information critical to the training of the model, such as algorithmic insights, training data, or training code.
>
> Consequently, Signatories commit to implementing state-of-the-art security mitigations designed to thwart such unauthorised access by well-resourced and motivated non-state-level adversaries, including insider threats from humans or AI systems, so as to meet at least the RAND SL3 security goal or equivalent, and achieve higher security goals (e.g. RAND SL4 or SL5) as specified in the Measures for this Commitment."

Google DeepMind's Frontier Safety Framework, page 5, relevant context:

> **Security mitigations against exfiltration risk are important for models reaching CCLs**. This is because the release of model weights may enable the removal of any safeguards trained into or deployed with the model, and hence access (including by threat actors) to any critical capabilities. Here, we **rely on the RAND framework to articulate the level of security recommended for each CCL**. When we reference RAND security levels, we are referring to the security principles in their framework, rather than the benchmarks (i.e. concrete measures) also described in the RAND report. Because AI security is an area of active research, we expect the concrete measures implemented to reach each level of security to evolve substantially

Page 6, at Machine Learning R&D autonomy level 1

> We **recommend security controls and detections at a level generally aligned with RAND SL 4 for this capability threshold**, but emphasize that this must be taken on by the frontier AI field as a whole.

G42's Frontier AI Safety Framework, page 9:

> **[Security Mitigation Levels] protect against the theft of model weights, model inversion, and sensitive data, as models reach higher levels of capability and risk**. Each tier customizes protections based on the assessed risk and capability of the model, ensuring G42's AI development remains both resilient and efficient, minimizing disruptions to functionality while maintaining robust security.

Microsoft's Frontier Governance Framework, page 8:

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

Anthropic's Responsible Scaling Policy, page 8-9, at 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).

We define “basic insider risk” as risk from an insider who does not have persistent or time-limited access to systems that process model weights. We define “sophisticated insider risk” as risk from an insider who has persistent access or can request time-limited access to systems that process model weights.

Anthropic's Responsible Scaling Policy, page 19, on the RSP v2.2 update:

excludes both sophisticated insiders and state-compromised insiders from the ASL- 3 Security Standard.

Page 9, list of criteria for ASL-3.

Measure II.7.1. General cybersecurity best practices

"Signatories shall implement general cybersecurity best practices so as to meet at least the RAND SL3 security goal. Illustrative examples of such best practices are:

(1) strong identity and access management practices; (2) strong protections against social engineering; (3) protection of wireless networks; (4) policies for untrusted removable media; (5) protection of premises from physical intrusion; or (6) regular software updates and patch management.

General cybersecurity best practices implemented pursuant to this Measure shall accord with: (1) up-to-date cybersecurity technical standards such as ISO/IEC 27001, NIST 800-53, SOC 2 and, if available, a harmonised standard on cybersecurity; and (2) other relevant Union law such as the NIS2 Directive and the Cyber Resilience Act to the extent applicable."

Magic's AGI Readiness Policy:

> 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.

Anthropic's Responsible Scaling Policy, page 8-9, at 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**.

Page 9, list of criteria for ASL-3.

> Security recommendations like Deploying AI Systems Securely (CISA/NSA/FBI/ASD/CCCS/GCSB/GCHQ), ISO 42001, CSA's AI Safety Initiative, and CoSAI; and standards frameworks like SSDF, **SOC 2**, **NIST 800-53**

G42's Frontier AI Safety Framework, page 9-10, at security level 3:

> Model weights and sensitive data are secured through thorough **Security Level 2 protocols**, as well as the following measures to ensure access to model weights is highly restricted: **multi-party and quorum-based approval for high-sensitivity operations, end-to-end encryption of model weights both at rest and in transit, automatic encryption key rotations at regular intervals**.
>
> Only **trusted users with verified credentials** are granted access to high-risk models. Access is role-based, aligned with user responsibilities, and supported by a zero-trust architecture to prevent unauthorized entry

Microsoft's Frontier Governance Framework, page 8:

> We expect scientific understanding of how to best secure the AI lifecycle will advance significantly in the coming months and years and will continue to contribute to, and apply, security best practices as relevant and appropriate. This includes **existing best practice defined in leading standards and frameworks**, such as **NIST SP 800-53**, NIST 800-218, **SOC 2**, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, and Deploying AI Systems Securely, as well as industry practices, including from the Frontier Model Forum.
>
> 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.

Google DeepMind's Frontier Safety Framework, page 5:

> **Security mitigations against exfiltration risk are important for models reaching CCLs**. This is because the release of model weights may enable the removal of any safeguards trained into or deployed with the model, and hence access (including by threat actors) to any critical capabilities. Here, we **rely on the RAND framework to articulate the level of security recommended for each CCL**. When we reference RAND security levels, we are referring to the security principles in their framework, rather than the benchmarks (i.e. concrete measures) also described in the RAND report. Because AI security is an area of active research, we expect the concrete measures implemented to reach each level of security to evolve substantially

> Page 6, at Machine Learning R&D autonomy level 1
>
> > We **recommend security controls and detections at a level generally aligned with RAND SL 4 for this capability threshold**, but emphasize that this must be taken on by the frontier AI field as a whole.

> Amazon's Frontier Model Safety Framework, page 6, Appendix A: **Amazon's Foundational Security Practices**:
>
> > Culture of security […]; Workforce vetting and careful monitoring for insider threats […]; Zero trust device and identity management […]; Strong separation of duties and least-privilege access […]; Physical security of datacenters […]; Hardware and software supply chain security […]; Specialized hardware designs to maximize security […]; Secure design, security reviews, and security testing […]; Secure development and deployment practices […]; Advanced IAM technologies that reduce external and internal risk […]; Advanced data encryption and key management capabilities […]; Access controls that eliminate or greatly limit operator access […]."

Measure II.7.2. Security assurance

"Signatories shall implement security measures to assure and test their security readiness against potential and actual adversaries so as to meet at least the RAND SL3 security goal. Illustrative examples of such security measures are:

(1) regular security reviews by an independent external party as necessary to mitigate systemic risks; (2) frequent red-teaming as necessary to mitigate systemic risks; (3) secure communication channels for third parties to report security issues; (4) competitive bug bounty programs to encourage public participation in security testing as necessary to mitigate systemic risks; (5) installation of Endpoint Detection and Response (EDR) and/or Intrusion Detection System (IDS) tools on all company devices and network components; or (6) the use of a security team to monitor for EDR alerts and perform security incident handling, response and recovery for security breaches in a timely manner.

Security measures implemented pursuant to this Measure shall accord with up-to-date security assurance technical standards, such as NIST 800-53 (CA-2(1), CA-8(2), RA-5(11), IR-4(14)), and NIST SP 800-115 5.2."

> Anthropic's Responsible Scaling Policy, page 9, list of criteria for ASL-3:
>
> > Security recommendations like Deploying AI Systems Securely (CISA/NSA/FBI/ASD/CCCS/GCSB /GCHQ), ISO 42001, CSA's AI Safety Initiative, and CoSAI; and standards frameworks like SSDF, **SOC 2**, **NIST 800-53**

> [...] **We will solicit both internal and external expert feedback on the** [**Safeguards Report**] as well as the CEO and RSO's conclusions to inform future refinements to our methodology. **For high-stakes issues, however, the CEO and RSO will likely solicit internal and external feedback on the report prior to making any decisions.**
>
> Page 10:
>
> 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

> G42's Frontier AI Safety Framework, page 10:
>
> **Internal and External Red Teaming**: **Rigorous testing by internal security teams**, supplemented by **external experts**, to identify weaknesses. Dynamic Threat Simulation and Response Testing: Regular adversarial simulations expose potential security weaknesses.

> Cohere's Secure AI Frontier Model Framework, page 9:
>
> We deploy a defense-in-depth strategy. This means we apply layered controls as part of our overall security management program across all systems and processes – for model development and general day-to-day operations – all of which directly contribute to how we secure our models and our internal environment. We **align our program to SOC 2 Type II and other recognized frameworks**, and we rigorously monitor the health and performance of our security controls throughout the year, performing real-time corrective action when needed. [...]
>
> Page 18:
>
> We are constantly improving our practices to better mitigate AI risks. To identify areas of improvement, **we engage with our customers, partners, and the wider public by:**
>
> - **Incentivizing third-party vulnerability discovery via clear protections for legitimate research practices in our Responsible Disclosure Policy**

> Microsoft's Frontier Governance Framework, page 8:
>
> We expect scientific understanding of how to best secure the AI lifecycle will advance significantly in the coming months and years and will continue to contribute to, and apply, security best practices as relevant and appropriate. This includes **existing best practice defined**

> **in leading standards and frameworks**, such as **NIST SP 800-53**, NIST 800-218, **SOC 2**, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, and Deploying AI Systems Securely, as well as industry practices, including from the Frontier Model Forum. [...]
>
> **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.

Measure II.7.3. Protection of unreleased model weights and associated assets

"Signatories shall implement security measures to protect unreleased model weights and associated assets so as to meet at least the RAND SL3 security goal. Illustrative examples of such security measures are:

(1) a secure internal registry of all devices and locations where model weights are stored; (2) access control and monitoring of access on all devices storing model weights, with alerts on copying to unmanaged devices; (3) storage on dedicated devices which host only data, code, and services treated with levels of sensitivity and security equivalent to those applied to the devices storing model weights and associated assets; (4) ensuring model weights are always encrypted in storage and transportation, in accordance with up-to-date best practice security standards, including encryption with at least 256-bit security and with encryption keys stored securely on a Trusted Platform Module (TPM) or in accordance with higher security standards that are established as best practice after the publication of the Code; (5) ensuring model weights are only decrypted for legitimate use to non-persistent memory; (6) implementing confidential computing if feasible in accordance with up-to-date industry standards, using hardware-based, and attested trusted execution environments to prevent unauthorised access to model weights while in use; or (7) restricting physical access to data centres and other sensitive working environments to required personnel only, along with regular inspections of such sites for unauthorised personnel or devices.

Security measures implemented pursuant to this Measure shall accord with up-to-date technical standards such as NIST 800-53."

> Amazon's Frontier Model Safety Framework, page 4:
>
> > **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**.

Page 6, Appendix A: Amazon's Foundational Security Practices

Anthropic's Responsible Scaling Policy, page 8-9, at 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**.

Page 9, list of criteria for ASL-3.

> Security recommendations like Deploying AI Systems Securely (CISA/NSA/FBI/ASD/CCCS/GCSB/GCHQ), ISO 42001, CSA's AI Safety Initiative, and CoSAI; and standards frameworks like SSDF, **SOC 2**, **NIST 800-53**

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**. [...]
>
> 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.

> G42's Frontier AI Safety Framework, page 9:
>
> **[Security Mitigation Levels] protect against the theft of model weights, model inversion, and sensitive data**, as models reach higher levels of capability and risk. Each tier customizes protections based on the assessed risk and capability of the model, ensuring G42's AI development remains both resilient and efficient, minimizing disruptions to functionality while maintaining robust security.

> xAI's Risk Management Framework (Draft), page 7:
>
> We intend to implement **appropriate information security standards sufficient to prevent Grok from being stolen** by a motivated non-state actor

Measure II.7.4. Interfaces and access control to unreleased model weights

"Signatories shall implement measures to harden interfaces and restrict access to unreleased model weights while in use so as to meet at least the RAND SL3 security goal. Illustrative examples of such interface and access control measures are:

(1) explicitly authorising only required software and persons for both direct or indirect access to model weights, enforced through multi-factor authentication mechanisms, and audited on a regular basis of at least every six months; (2) thorough review by a security team of any software interface accessing model weights for vulnerabilities or data leakage; or (3) hardening interfaces with access to the model weights to reduce the risk of data and weight exfiltration, using methods such as output rate limiting.

Interface and access control measures implemented pursuant to this Measure shall accord with up-to-date technical standards such as NIST SP 800-171, INCITS 359-2004, and NIST 800-53."

> Microsoft's Frontier Governance Framework, page 8:
>
> We expect scientific understanding of how to best secure the AI lifecycle will advance significantly in the coming months and years and will continue to contribute to, and apply,

> security best practices as relevant and appropriate. This includes existing best practice defined in **leading standards and frameworks, such as NIST SP 800-53, NIST 800-218, SOC 2**, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, and Deploying AI Systems Securely, as well as industry practices, including from the Frontier Model Forum. [...]
>
> 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. [...]
>
> 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.

Anthropic's Responsible Scaling Policy, page 9, list of criteria for ASL-3:

> Building **strong perimeters and access controls** around sensitive assets to ensure AI models and critical systems are protected from unauthorized access. We expect this will include a combination of **physical security, encryption, cloud security, infrastructure policy, access management, and weight access minimization and monitoring**. [...]
>
> Security recommendations like Deploying AI Systems Securely (CISA/NSA/FBI/ASD/CCCS/GCSB /GCHQ), ISO 42001, CSA's AI Safety Initiative, and CoSAI; and standards frameworks like SSDF, SOC 2, NIST 800-53.

Page 15, ASL-2 standard in Appendix:

> Features like **zero trust architecture** should require access from approved devices. Strict protocols must be deployed to regulate weight copies on company networks and limit storage to only approved, restricted systems

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

> **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).

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**.

> G42's Frontier AI Safety Framework, page 10-11:
>
> > Only **trusted users with verified credentials** are granted access to high-risk models. Access is **role-based, aligned with user responsibilities**, and supported by a **zero-trust architecture** to prevent unauthorized entry

Measure II.7.5. Insider threats

"Signatories shall implement measures to screen for and protect against insider threats so as to meet at least the RAND SL3 security goal, including threats in the form of self-exfiltration or sabotage carried out by GPAISRs. Illustrative examples of such measures are:

(1) background checks on employees and contractors that have or might reasonably obtain access to unreleased model weights of their models, associated assets, or systems that control the storage or use of such unreleased model weights; (2) the provision of training on recognising and reporting insider threats; or (3) sandboxes around GPAISRs.

Measures implemented to screen for and protect against insider threats pursuant to this Measure shall accord with up-to-date technical standards such as NIST 800-53 (PM-12, PS-3)."

> Anthropic's Responsible Scaling Policy, page 15:
> [For ASL-2]
>
> > Workforce: People-critical processes must represent a key aspect of cybersecurity. **Mandatory periodic infosec training educates all employees on secure practices**, like proper system configurations and strong passwords, and fosters a proactive "security mindset." Fundamental infrastructure and policies promoting secure-by-design and secure-by-default principles should be incorporated into the engineering process. **An insider risk program should tie access to job roles.** Rapid incident response protocols must be deployed.
>
> Page 9:
> [For ASL-3]
>
> > **We will implement robust insider risk controls to mitigate most insider risk, but consider mitigating risks from highly sophisticated state-compromised insiders to be out of scope for ASL-3**. We are committed to further enhancing these protections as a part of our ASL-4

> preparations.

> Microsoft's Frontier Governance Framework, 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**.

> Amazon's Frontier Model Safety Framework, page 5:
>
> Focused monitoring of threats and abuse by Amazon threat teams. Amazon's threat intelligence, Trust & Safety, and **insider threat teams are building additional capabilities to track advanced threat actors and how they interact with and attempt to subvert security measures surrounding AI models**. Learnings from this observation and analysis will provide important signals about where additional capabilities and layers of defense are needed to protect against any kind of "side door" or "back door" access to frontier models.

Measure II.7.6. Regime of applicability

"Signatories shall ensure that the security measures set out in Measures II.7.1–II.7.5: (1) apply along the entire model lifecycle from before training until secure deletion or a decision to release the model weights or associated assets; and (2) without prejudice to (1), prioritise the implementation of security measures in accordance with the systemic risk stemming from the GPAISR."

> Anthropic's Responsible Scaling Policy, page 2:
>
> Security Standards: Security Standards are technical, operational, and policy measures to protect AI models–particularly their weights and associated systems–from unauthorized access, theft, or compromise by malicious actors. **Security Standards are intended to maintain the integrity and controlled use of AI models throughout their lifecycle, from development to deployment.**

> Microsoft's Frontier Governance Framework, page 2:
>
> Targeted and proportional. The framework monitors Microsoft's most capable AI models for leading indicators of high-risk capabilities and triggers deeper assessment if leading indicators are observed.

> Page 6:
>
> > 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:
>
> > **Defense in depth across the lifecycle**, applying multiple layers of security controls that provide redundancy in case some controls fail. Model weights are encrypted.

> Google DeepMind's Frontier Safety Framework, page 4:
>
> > For security, we have several levels of mitigations, allowing **calibration of the robustness of security measures to the risks posed**.

> OpenAI's Preparedness Framework v1, page 2:
>
> > **Security is appropriately tailored to any model that has a "high" or "critical" pre-mitigation level of risk** (as defined in the Scorecard below) to prevent model exfiltration.

> Cohere's Secure AI Frontier Model Framework, pages 8-9:
>
> > Cohere's security-first culture drives how we work together to design, operate, continuously monitor, and secure both our internal environment. [...]
> > This means we **apply layered controls as part of our overall security management program across all systems and processes – for model development and general day-to-day operations** – all of which directly contribute to how we secure our models and our internal environment.

Measure II.7.7. Limited disclosure

> "Signatories shall ensure that any publicly available copy of the Framework or the Model Report discloses security mitigations considered and/or implemented pursuant to Measures II.7.1–II.7.5: (1) with the greatest level of detail possible given the state of science; (2) without undermining their effectiveness; and (3) with redactions as necessary to prevent substantially increased systemic risk that could result from, e.g., undermining the effectiveness of technical safety mitigations (in accordance

with Commitment II.6) or divulging sensitive commercial information to a degree disproportionate to the societal benefit of the publication."

> xAI's Risk Management Framework (Draft), page 6:
>
> For transparency and third-party review, we may publish the following types of information listed below. **However, to protect public safety, national security, and our intellectual property, we may redact information from our publications.** We may provide relevant and qualified external red teams or relevant government agencies unredacted versions.

> Anthropic's Responsible Scaling Policy, page 12:
>
> 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.

> Microsoft's Frontier Governance Framework, page 10:
>
> We also highlight the value of learning from experts outside of AI, including those with expertise in measurement science and in scientific domains like chemistry and biology, as well as those with knowledge of managing the risks of other complex technologies. **We will continue to share information and lessons from our governance efforts, taking care to do so in a way that minimizes the chance of inadvertently increasing risk, including by being judicious about sharing sensitive safety- and security-relevant information.**

> OpenAI's Preparedness Framework v1, page 24:
>
> Internal visibility: The Preparedness Framework, reports and decisions will be documented and visible to the BoD and within OpenAI (**with redactions as needed given internal compartmentalization of research work**). This also includes any audit trails created from the below.

Measure II.7.8. Higher security goals

"Signatories shall advance research on and implement more stringent security mitigations, in line with at least the RAND SL4 security goal or equivalent, if they reasonably foresee security threats from RAND OC4-level adversaries."

> Microsoft's Frontier Governance Framework, page 7:
>
> > **We expect scientific understanding of how to best secure the AI lifecycle will advance significantly in the coming months and years and will continue to contribute to, and apply, security best practices as relevant and appropriate**. This includes existing best practice defined in leading standards and frameworks, such as NIST SP 800-53, NIST 800-218, SOC 2, Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models, and Deploying AI Systems Securely, as well as industry practices, including from the Frontier Model Forum. Security safeguards are scaled up depending on the model's pre-mitigation scores, with more robust measures applied to models with high and critical risk levels.

> OpenAI's Preparedness Framework v1, page 3:
>
> > Finally, OpenAI's primary fiduciary duty is to investment in improving alignment humanity, and we are committed to doing the research required to make AGI safe. Therefore, the Preparedness Framework is meant to be just one piece of our overall approach to safety and alignment, mitigating bias, hallucination, and misuse , facilitating which also includes democratic inputs to AI, methods, **investing significantly in security voluntary commitments and safety research**. This is also one more way in which we are meeting our and trust in AI that we made in July 2023.

> Anthropic's Responsible Scaling Policy, page 4:
>
> > At minimum, the **ASL-4 Security Standard (which would protect against model-weight theft by state-level adversaries) is required**, although we expect a higher security standard may be required. As with AI R&D-4, we also expect an affirmative case will be required.

> Google DeepMind's Frontier Safety Framework, page 6:
>
> > We recommend security controls and detections **at a level generally aligned with RAND SL 4 for [Machine Learning R&D autonomy level 1] capability threshold**, but emphasize that this must be taken on by the frontier AI field as a whole.

> G42's Frontier AI Safety Framework, page 11:
>
> > Level 4 [objectives]: Security strong enough to resist concerted attempts with support from state programs to steal model weights.

Commitment II.8. Safety and Security Model Reports

"Signatories commit to reporting to the AI Office about their implementation of the Code, and especially the application of their Framework to the development, making available on the market, and/or use of their GPAISRs, by creating a Safety and Security Model Report (hereafter, a "Model Report") for each GPAISR which they make available on the market, which will document, as specified in the Measures for this Commitment: (1) the results of systemic risk assessment and mitigation for the model in question; and (2) justifications of decisions to make the model in question available on the market."

> Anthropic's System Card: Claude Opus 4 & Claude Sonnet 4, page 8:
>
> > We **document all evaluation results and risk assessments** to maintain transparency and enable continuous improvement of our safety processes.

> Anthropic's Responsible Scaling Policy, executive summary:
>
> > To advance the public dialogue on the regulation of frontier AI model risks and to enable examination of our actions, we will also **publicly release key materials related to the evaluation and deployment of our models** with sensitive information removed and solicit input from external experts in relevant domains.
>
> Pages 12-13, at 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.

> Meta's Frontier AI Framework, page 9, at Transparency:
>
> > We also plan to continue **sharing relevant information about how we develop and evaluate our models responsibly**, including through artefacts like model cards and research papers.

> Microsoft's Frontier Governance Framework, page 9:
>
> > **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.

> Page 8:
>
> > [We will utilize] 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. This documentation will also include a **summary of evaluation results, the deeper capability assessment, and safety and security mitigations**.

> OpenAI's Preparedness Framework v2, page 12, at 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 12 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.

Measure II.8.1. Level of detail

"Signatories shall provide a level of detail in the Model Report that:

(1) is proportionate to the level of systemic risk that the model in question reaches, or that the Signatory reasonably foresees that model to reach along the entire model lifecycle; and

(2) allows the AI Office to understand how a Signatory has implemented its systemic risk assessment and mitigation measures pursuant to the Code.

A Model Report shall contain sufficient reasoning as to why (1) and (2) are satisfied."

> Anthropic's Claude 3.7 System Card, page 6:
>
> > Because the initial evaluation results revealed complex patterns in model capabilities, we supplemented our standard process with multiple rounds of feedback between FRT [Frontier Red Team] and AST [Alignment Stress Testing]. The teams worked iteratively, continually refining their respective analyses and challenging each other's assumptions to reach a thorough understanding of the

> model's capabilities and their implications. **This more comprehensive process reflected the complexity of assessing a model with increased capabilities relevant to the capability threshold.**

NVIDIA's Frontier AI Risk Assessment, page 6:

> The MR (Model Risk) score is correlated to the maximum permissible harm relative to our trustworthy AI principles. **High risk models require more intensive scrutiny,** increased oversight and face stricter development and operational constraints.

Anthropic's Responsible Scaling Policy, Executive Summary, at Capability assessment:

> We will routinely test models to determine whether their capabilities fall sufficiently far below the Capability Thresholds such that the ASL-2 Standard remains appropriate. **We will first conduct preliminary assessments to determine whether a more comprehensive evaluation is needed. For models requiring comprehensive testing, we will assess whether the model is unlikely to reach any relevant Capability Thresholds absent surprising advances in widely accessible post-training enhancements**. If, after the comprehensive testing, we determine that the model is sufficiently below the relevant Capability Thresholds, then we will continue to apply the ASL-2 Standard. If, however, we are unable to make the required showing, we will act as though the model has surpassed the Capability Threshold. This means that we will both upgrade to the ASL-3 Required Safeguards and **conduct a follow-up capability assessment to confirm that the ASL-4 Standard is not necessary.**

Google DeepMind's Frontier Safety Framework, page 3:

> We also intend to evaluate any of [the] models that could indicate an exceptional increase in capabilities over previous models, and where appropriate, assess the likelihood of such capabilities and risks before and during training.
>
> **To do so, we will define a set of evaluations called "early warning evaluations," with a specific "alert threshold" that flags when a CCL [Critical Capability Threshold] may be reached before the evaluations are run again. [...]** Where necessary, **early warning evaluations may be supplemented by other evaluations to better understand model capabilities** relative to our CCLs.

Measure II.8.2. Reasons for deciding to make the model in question available on the market

"Signatories shall provide in the Model Report the reasoning and information used to justify a decision to make the model in question available on the market including for the first time, by stating:

(1) a clear chain of reasoning, based on the information and data provided in the Model Report, which provides sufficient justification that the systemic risk of the GPAISR is acceptable as per Measure II.5.1, including a sufficient safety margin pursuant to Measure II.4.12 and an explanation thereof to evaluate it;

(2) the conditions under which the Signatory's reasoning in point (1) would no longer hold; and

(3) whether and how independent external assessors informed a decision to make the model in question available on the market or use it, such as through the evaluation of models pursuant to Commitment II.11."

> OpenAI's Preparedness Framework v2, pages 10-11:
>
> We compile the information on the planned safeguards needed to minimize the risk of severe harm into a **Safeguards Report. The Safeguards Report should include the following information:**
> - Identified ways a risk of severe harm can be realized for the given deployment, each mapped to the associated security controls and safeguards
> - Details about the efficacy of those safeguards
> - An assessment on the residual risk of severe harm based on the deployment
> - Any notable limitations with the information provided
>
> [...] **Based on this [the Safeguards Report], SAG [Safety Advisory Group] then has the following decision points:**
> 1. **SAG can find that it is confident that the safeguards sufficiently minimize the associated risk of severe harm for the proposed deployment, and recommend deployment.** [...]
>
> Page 16:
>
> Safeguards should sufficiently minimize the risk of severe harm associated with misuse of the model's capabilities. **This can be done by establishing that all plausible known vectors of enabling severe harm are sufficiently addressed by one or more of the following claims:**
>
> - Robustness: Users cannot use the model to cause the harm because they cannot elicit the capability, such as because the model is modified to refuse to provide assistance to harmful tasks and is robust to jailbreaks that would circumvent those refusals.
> - Usage Monitoring: If a model does not refuse and provides assistance to harmful tasks, monitors can stop or catch adversarial users before their misuse has achieved an unacceptable scale of harm, through a combination of automated and human detection and enforcement within an acceptable time frame.
> - Trust-based Access: The actors who gain access to the model will not use it in a

> way that presents an associated risk of severe harm under our threat model. The table below illustrates possible examples of safeguards and safeguard efficacy assessments we could consider to claim we have sufficiently minimized the risks of severe harm associated with a High level of capability under any of our Tracked Categories.

Pages 12-13:

> 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).
>
> [...] **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.

Anthropic's Claude 3.7 System Card, pages 5-6:

> Our **release decision process is guided by our Responsible Scaling Policy (RSP),** which provides a framework for evaluating and managing potential risks associated with increasingly capable AI systems. [...]
>
> For each domain, we conduct extensive testing to determine the ASL of the safeguards required. [...] The **final release decision requires verification that safety measures appropriate to the ASL level are in place, including monitoring systems and incident response protocols.** We document all evaluation results and risk assessments to maintain transparency and enable continuous improvement of our safety processes.

Pages 23-24, CBRN evaluations thresholds:

> To test whether the model causes uplift, we evaluate whether models provide both the sufficient knowledge and skills assistance required to acquire and misuse CBRN weapons. Our evaluations include automated knowledge evaluations, automated skill-testing questions, uplift studies designed to proxy real world tasks, external red teaming by experts, and long-form task-based agentic evaluations.

Page 25, Biological Risks Threshold:

> We pre-registered an 80% average score in the uplifted group as indicating ASL-3 capabilities, with scores below 50% suggesting multiple key steps were incorrect. In addition, further threat analysis sug- gested that achieving ≥ 5X total uplift in a "real-world" uplift trial would result in significant additional risk, while ≤ 2.8X uplift would bound risk to an acceptable level.

Page 24:

> We **provided a copy of our capabilities report to [the U.S. and U.K. AISIs) and to METR for feedback**.

Anthropic's Responsible Scaling Policy, page 6:

> We will **compile a Capability Report that documents the findings from the comprehensive assessment, makes an affirmative case for why the Capability Threshold is sufficiently far away, and advances recommendations on deployment decisions.**
>
> - The report will be escalated to the CEO and the Responsible Scaling Officer [RSO], who will (1) make the ultimate determination as to whether we have sufficiently established that we are unlikely to reach the Capability Threshold and (2) decide any deployment-related issues.
> - In general, as noted in Sections 7.1.4 and 7.2.2, we will **solicit both internal and external expert feedback on the report** as well as the CEO and RSO's conclusions to inform future refinements to our methodology. For high-stakes issues, however, the CEO and RSO will likely solicit internal and external feedback on the report prior to making any decisions.
> - If the CEO and RSO decide to proceed with deployment, they will **share their decision–as well as the underlying Capability Report, internal feedback, and any external feedback–with the Board of Directors and the Long-Term Benefit Trust before moving forward.**

Page 9:

> We will compile a Safeguards Report for each Required Safeguard that documents our implementation of the [ASL-3 Required Safeguards] above, **makes an affirmative case for why we have satisfied them,** and advances recommendations on deployment decisions.

Page 13:

> We may also solicit external expert input prior to making final decisions on the capability and safeguards assessments.

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.

Page 6, at Holistic risk assessment:

> The results of capability evaluation and an assessment of risk factors external to the model then inform a determination as to whether a model has a tracked capability and to what level. **This includes assessing the impact of potential system-level mitigations and societal and institutional factors that can impact whether and how a hazard materializes. This holistic risk assessment also considers the marginal capability uplift a model may provide over and above currently available tools and information, including currently available open-weights models.**

Page 9:

> 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. This documentation will also include a summary of evaluation results, the deeper capability assessment, and safety and security mitigations.

Google DeepMind's Gemini 2.5 System Card, pages 34-35:

> Gemini 2.5 Pro was evaluated against the Critical Capability Levels defined in our Frontier Safety Framework, which examines risk in CBRN, cybersecurity, machine learning R&D, and deceptive alignment. **Based on these results, we find that Gemini 2.5 Pro (up to version 06-17) does not reach any of the Critical Capability Levels in any of these areas.**
>
> However, it showed some ability in all four areas. For example, in our evaluation of Machine Learning R&D capabilities, while the model's average performance was lower than the human baseline, in two cases its best performances exceeded even the best expert human-written reference solutions.
>
> Gemini 2.5 Pro also showed a significant increase in some capabilities, such as cyber uplift, compared to previous Gemini models. Following our Frontier Safety Framework, we are putting in place a response plan, including conducting higher frequency testing and accelerating mitigations for the Cyber Uplift Level 1 CCL. As reported above, no model reached the CCL in these additional tests.

Google DeepMind's Frontier Safety Framework, page 4:

> 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. This includes, as appropriate, safety fine-tuning, misuse filtering and detection, and response protocols.
>
> b. Assessing the robustness of these mitigations against the risk posed through assurance evaluations and threat modeling research. The assessment takes the form of a safety case, taking into account factors such as the likelihood and consequences of misuse.
>
> 2. *Pre-deployment review of safety case:* **general availability deployment of a model takes place only after the appropriate corporate governance body determines the safety case regarding each CCL the model has reached to be adequate.**
>
> 3. *Post-deployment review of safety case:* the safety case will be updated through red-teaming and revisions to our threat models. The safeguards for the model may be updated as well to ensure continued adequacy.

Measure II.8.3. Documentation of systemic risk assessment and mitigation

"In the Model Report,[8] Signatories shall:

(1) document and explain their systemic risk selection as per Measure II.3.1, including a description of their efforts to select other systemic risks that are specific to the high-impact capabilities of the GPAISR, with further reference to Appendices 1.2, 1.3, and 1.4 where appropriate;

> OpenAI's Preparedness Framework v1, page 23, at Operations:
>
> The Preparedness team is responsible for:
>
> I. maintaining and updating the Scorecard, including designing and running evaluations to provide Scorecard inputs and collecting relevant information on monitored misuse, red-teaming, and intelligence
>
> **II. monitoring for unknown unknowns and making the case for inclusion in the Preparedness Framework of any new risk categories as they emerge**
>
> III. ensuring the risk level distinctions in the Tracked Risk Categories section are appropriate given developments in frontier AI models, and suggesting updates to these levels if needed
>
> IV. forecasting potential changes to catastrophic risk levels, and summarizing evidence for an "early warning" / "heads up" as needed
>
> **V. providing a monthly report (sent to the SAG [Safety Advisory Group], Leadership and BoD [Board of Directors]) synthesizing the above with any**

[8] The quotes below are sometimes taken from companies' Safety Frameworks, rather than Model Reports. However, we still consider this evidence of industry precedent, as it stands as evidence that the process of producing this documentation is conducted, even if it is not publicly published.

> **potential protective actions (the SAG Chair, OpenAI Leadership, and/or BoD can adjust this cadence as needed**

OpenAI's Preparedness Framework v2, page 4:

> **We evaluate whether frontier capabilities create a risk of severe harm through a holistic risk assessment process.** This process draws on our own internal research and signals, and where appropriate incorporates feedback from academic researchers, independent domain experts, industry bodies such as the Frontier Model Forum, and the U.S. government and its partners, as well as relevant legal and policy mandates.
>
> Where we determine that a capability presents a real risk of severe harm, we may decide to monitor it as a Tracked Category or a Research Category.
>
> Tracked Categories are those capabilities which we track most closely, measuring them during each covered deployment and preparing safeguards for when a threshold level is crossed. **We treat a frontier capability as a Tracked Category if the capability creates a risk that meets five criteria:**
>
> 1. Plausible: It must be possible to identify a causal pathway for a severe harm in the capability area, enabled by frontier AI.
> 2. Measurable: We can construct or adopt capability evaluations that measure capabilities that closely track the potential for the severe harm.
> 3. Severe: There is a plausible threat model within the capability area that would create severe harm.
> 4. Net new: The outcome cannot currently be realized as described (including at that scale, by that threat actor, or for that cost) with existing tools and resources (e.g., available as of 2021) but without access to frontier AI.
> 5. Instantaneous or irremediable: The outcome is such that once realized, its severe harms are immediately felt, or are inevitable due to a lack of feasible measures to remediate.
>
> **We review and update Tracked Categories periodically or when we learn significant new information.**

Microsoft's Frontier Governance Framework, page 3:

> This framework tracks the **following capabilities that we believe could emerge over the short-to-medium term and threaten national security or pose at-scale public safety risks if not appropriately mitigated.** In formulating this list, we have benefited from the advice of both internal and external experts. [...] AI technology continues to develop rapidly, and there remains uncertainty over which capabilities may emerge and when. We continue to study a range of potential capability- related risks that could emerge, conducting ongoing assessment of the severity and likelihood of these risks. We then operationalize the highest-priority risks through this framework. **We will revisit our list of tracked capabilities frequently, ensuring it remains up to date in light of technological developments and improved understanding of model capabilities, risks, and**

> **mitigations**.

> G42 Frontier AI Safety Framework, page 4:
>
> > **To produce this list [of frontier capability thresholds], G42 both conducted our own internal risk analysis and received input from external AI safety experts.** Initially G42 identified potential capabilities across several domains, including biological risks, cybersecurity, and autonomous operations in specialized fields. We then collaborated with METR and SaferAI to refine our list, prioritizing capabilities based on their potential impact and how feasibly they can be measured and monitored. [...] **We plan to integrate decisions on whether to expand our monitoring to include additional hazardous capabilities into our regular framework review process. This includes both our scheduled internal reviews and our annual reviews by third parties. In making these decisions, we expect to consider factors such as: "near miss" incidents, whether internal or industry-wide; recommendations from trusted external experts; as well as changes in industry standards for AI risk management.**

> Anthropic's Responsible Scaling Policy, page 5:
>
> > For models requiring comprehensive testing, we will assess whether the model is unlikely to reach any relevant Capability Thresholds absent surprising advances in widely accessible post-training enhancements. To make the required showing, we will need to satisfy the following criteria:
> >
> > *Threat model mapping:* **For each capability threshold, make a compelling case that we have mapped out the most likely and consequential threat models:** combinations of actors (if relevant), attack pathways, model capability bottlenecks, and types of harms. **We also make a compelling case that there does not exist a threat model that we are not evaluating that represents a substantial amount of risk.**
>
> Page 4:
>
> > 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.

(2) document and explain the systemic risk estimation process performed (pursuant to Measure II.4.1), and provide a justification of all decisions made as part of that process, in particular the choice of qualitative and/or quantitative estimates, and how the estimates relate to systemic risk tiers;

> Anthropic's System Card: Claude Opus 4 & Claude Sonnet 4, page 93:
>
> > Our threat analysis indicated that total uplift $\geq 5\times$ (or raw uplift ≥ 0.8) would create significant additional risk, while uplift $\leq 2.8\times$ would keep risk at acceptable levels. Text-based uplift trials are substantially weaker proxies for real-world scenarios—which

> involve additional factors like tacit knowledge, materials access, and actor persistence—but we nevertheless adopted a similar uplift threshold to be conservative in our ASL-3 rule-out decisions.

OpenAI's o3 and o4-mini System Card, page 15:

> **Neither OpenAI o3 nor o4-mini met the High threshold for cybersecurity.** Neither model was able to sufficiently succeed in the professional-level Capture the Flag challenges, and neither model was able to solve real-world-relevant range scenarios without being explicitly given solver code. We do not consider them as posing a high autonomous operations risk, but note that both o3 and o4-mini show more capabilities in executing autonomous cyber operations tasks than previously released models.

Page 21:

> OpenAI o3 and o4-mini demonstrate improved performance on software engineering and AI research tasks relevant to AI self-improvement risks. In particular, their performance on SWE-Bench Verified demonstrates the ability to competently execute well-specified coding tasks. **However, these tasks are much simpler than the work of a competent autonomous research assistant; on evaluations designed to test more real-world or open-ended tasks, the models perform poorly, suggesting they lack the capabilities required for a High classification.**

The report gives tables justifying why evaluations track the risk categories in the Preparedness Framework v2. For instance, following table shows justification for why the evaluations chosen track the risk category of AI Self-Improvement, on page 22:

Table 15: Overview of AI Self-Improvement evaluations

Evaluation	Capability	Description
OpenAI Research Engineer Interview: Multiple Choice and Coding	Basic short horizon ML expertise	How do models perform on 97 multiple choice questions derived from OpenAI ML interview topics? How do models perform on 18 self-contained coding problems that match problems given in OpenAI interviews?
SWE-bench Verified	Real-world software engineering tasks	Can models resolve GitHub issues, given just a code repository and issue description?
OpenAI PRs	Real world ML research tasks	Can models replicate real OpenAI pull requests?
SWE-Lancer	Real world software engineering tasks	How do models perform on real-world, economically valuable full-stack software engineering tasks?
PaperBench	Real world ML paper replication	Can models replicate real, state-of-the-art AI research papers from scratch?

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.**
>
> Page 6:
>
> The results of the deeper capability assessment are used to assign a model a pre-mitigation score of low, medium, high, or critical for each tracked capability on the basis of the framework's capability thresholds (see Appendix I). Models assessed as posing low or medium risk may be deployed with appropriate safeguards as outlined below. Models assessed as having a high or critical risk are subject to further review and safety and security mitigations prior to deployment.

(3) where they make use of the "safely derived model" concept according to Measure II.4.2, provide a justification of how the criteria for a "safe originator model" (as per Measure II.4.2, point (1)) and the criteria for the "safely derived model" (as per Measure II.4.2, point (2) and point (3)) are fulfilled;

> OpenAI's o3-mini System Card, page 2:
>
> o3-mini includes small incremental post training improvements upon o3-mini-near-final-checkpoint, though the base model is the same. We **determined that risk recommendations based on red teaming and the two Persuasion human eval results conducted on the o3-mini-near-final-checkpoint remain valid** for the final release checkpoint.

> G42's Safety Framework, page 5:
>
> When G42 fine-tunes or adapts an existing open-source model, our assessments of whether our models meet Frontier Capability Thresholds may include relevant evaluation results for the original open-source models when analyzing the impact of our modifications. **If we determine that the open-source model has already undergone high-quality assessments for biological threats and offensive cybersecurity capabilities, and our modifications do not enhance its capabilities in these risk areas, we may place greater reliance on the original evaluation results.**

(4) document and explain the limitations of and uncertainties about the systemic risk assessment carried out for the model in question, including limitations of the model evaluation methods used (pursuant to Measure II.4.5) and the level of uncertainty in their model evaluation results, and particularly: (a) safety margins (pursuant to Measure II.4.12); (b) any lower levels of rigour in model evaluation adopted; (c) the reasons justifying such lower levels; and (d) the consequences for the model evaluation conducted;

OpenAI's GPT-4.5 System Card, page 8:

> We ran automated preparedness evaluations throughout training and on early post-trained checkpoints of GPT-4.5, as well as a final automated eval sweep on the launched model. For the evaluations below, we also tested a variety of elicitation methods, including custom scaffolding and prompting where relevant. However, **Preparedness evaluations represent a lower bound for potential capabilities;** additional prompting or fine-tuning, longer rollouts, novel interactions, or different forms of scaffolding could elicit behaviors beyond what we observed in our tests or the tests of our third-party partners.
>
> We calculate 95% confidence intervals for pass@1 using the standard bootstrap procedure that resamples model attempts per problem to approximate the metric's distribution. **While widely used, this method can underestimate uncertainty for very small datasets, as it captures only sampling variance (randomness in the model's performance on the same problems across multiple attempts) rather than all problem-level variance (variation in problem difficulty or pass rates). This can lead to overly tight confidence intervals, especially when a problem's pass rate is near 0% or 100% with few attempts.** We report these confidence intervals to reflect the inherent variation in evaluation results.

Page 7:

> **Capability evaluations after a model has been fully trained only allow third parties to make limited safety assurances.** For example, testing models during development, testing models for sandbagging, or accounting for known elicitation gaps may be important for robust safety assurances.

OpenAI's o3-mini System Card, page 8:

> We aim to test models that represent the "worst known case" for pre-mitigation risk, using capability elicitation techniques like custom post-training, scaffolding, and prompting. **However, our evaluations should still be seen as a lower bound for potential risks.** Additional prompting or fine-tuning, longer rollouts, novel interactions, or different forms of scaffolding are likely to elicit behaviors beyond what we observed in our tests or the tests of our third-party partners. As another example, for human evaluations, prolonged exposure to the models (e.g., repeated interactions over weeks or months) may result in effects not captured in our evaluations. **Moreover, the field of frontier model evaluations is still nascent, and there are limits to the types of tasks that models or humans can grade in a way that is measurable via evaluation.** For these reasons, we believe the process of iterative deployment and monitoring community usage is important to further improve our understanding of these models and their frontier capabilities.
>
> We show these confidence intervals to convey eval variance, but as always, **please note that all of our evaluation results can only be treated as a lower bound of potential model capability, and that additional scaffolding or improved capability elicitation could substantially increase observed performance.**

The following was mentioned multiple times in the report, on pages 8, 11, 18, 21, 25, 31:

> **As always, we note that all eval results likely represent lower bounds on model**

> **capability, because additional scaffolding or improved capability elicitation could substantially increase observed performance.**

> OpenAI's o3 and o4-mini System Card, page 11:
>
> We calculate 95% confidence intervals for pass@1 using the standard bootstrap procedure that resamples model attempts per problem to approximate the metric's distribution. While widely used, this method can underestimate uncertainty for very small datasets, as it captures only sampling variance (randomness in the model's performance on the same problems across multiple attempts) rather than all problem-level variance (variation in problem difficulty or pass rates). **This can lead to overly tight confidence intervals, especially when a problem's pass rate is near 0% or 100% with few attempts. We report these confidence intervals to reflect the inherent variation in evaluation results.**
>
> o3 and o4-mini's ability to browse the internet creates challenges for evaluating the model's capabilities. In many Preparedness evaluations, we aim to understand the model's ability to reason or solve problems. If the model can retrieve answers from the internet, then it may provide solutions without working through the problems itself, and could receive a high score without actually demonstrating the capability that the evaluation is intended to measure. **In this situation, the score would be artificially elevated and would be a poor measure of the model's true capability, a problem known as "contamination" of the evaluation. [...] We note in the interpretations for each evaluation where evaluations are known or suspected to be contaminated. Where relevant, we only plot no-browsing models.**

> OpenAI's o1 system Card, page 14:
>
> METR does not consider this result [human-level performance with tailored scaffolding] a good upper bound on what could be achieved with additional scaffolding improvements, finetuning or other post-training enhancements.

> Microsoft's Frontier Governance Framework, page 6:
>
> Evaluations are documented in a consistent fashion setting out the capability being evaluated, the method used, and evaluation results. **This evaluation also includes a statement on the robustness of the evaluation method used and any concerns about the effectiveness or validity of the evaluation.**

> Anthropic's Claude 3 Model Card, page 25:
>
> **Based on conversations with global experts, it is difficult to define strict pass/fail criteria for ASL-3 misuse evaluations with high confidence. Instead, we set the bar relatively low, such that passing the misuse evaluations would trigger discussion with relevant experts and extensive transcript reviewing to determine whether the model presents a true risk or the thresholds are too conservative.** The model is considered to pass the early indicator threshold if either of the following occur:
>
> - Sub-expert-level individuals achieve a greater than 25% increase in accuracy on a set of advanced bioweapon-relevant questions when using the model

> compared to using Google alone.
>
> - The model exhibits a 25% jump on one of two biological question sets when compared to the Claude 2.1 model. These tests are (1) a multiple choice question set on harmful biological knowledge and (2) a set of questions about viral design.

(5) document and explain, for each model evaluation used to assess systemic risks of their GPAISRs, their model elicitation work (pursuant to Measure II.4.6);

> Anthropic's Claude 3.7 System Card describes the elicitation process for each of its evaluations (pages 23, 29, 34). For instance, quoting from page 34):
>
> > **The model has access to a tool harness for code manipulation, including Bash, Python, and Code Editor capabilities. We leverage chain-of-thought reasoning and scaffolded reflection to enhance context utilization.** A key feature is our new Terminal Tool, enabling asynchronous management of multiple terminal sessions. The model operates within a Kali-based environment equipped with standard penetration testing tools like pwntools and metasploit. For evaluation purposes, we tested both the production model and a helpful-only version of Claude 3.7 Sonnet to ensure results aren't affected by safety mitigations we use in production.

> OpenAI's o3-mini System Card, page 7:
>
> > For the evaluations below, **we tested a variety of methods to best elicit capabilities in a given category, including custom model training, scaffolding, and prompting where relevant.**

> Anthropic's Responsible Scaling Policy, page 6:
>
> > [Evaluations should] 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 o3 and o4-mini System Card, page 11:
>
> > For the evaluations below, we also tested a variety of elicitation methods, including custom post-training (e.g., to create a "helpful-only" model), scaffolding, and prompting where relevant.

(6) document and explain the ways in which they have taken into account AI systems information pursuant to Measure II.4.7 for each model evaluation performed and explain why this was appropriate for an effective assessment of the systemic risk in question;

> OpenAI's o1-preview System Card, page 30:
>
> > **Browser:** To elicit capabilities, we work with Ranger [an external testing company], which developed a browsing harness that provides the model preprocessed HTML (with optional screenshots) and asks the model for the next action to take. **We find that using an external scaffold enables us to measure capabilities in real-world deployments.**

> 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.

> NVIDIA's Frontier AI Risk Assessment, page 10:
>
> > The ODD [operational design domain] refers to the specific conditions under which a product is intended to operate effectively and includes a description of the impacted stakeholders. **The ODD can therefore be used to help structure a suitable test dataset for model evaluation.**
>
> Page 4:
>
> > When developing an AI model, **it is important to record assumptions about the intended use case (if any) to provide context around model quality and any known limitations.** The output from these assessments are documented in our model cards and **supports our customers when safely integrating our models into downstream applications or systems.**

(7) document for each systemic risk assessed: (a) the qualifications of, as well as the level of access, resources, and support provided to, their internal model evaluation teams (as required by Measure II.4.11); and (b) the same information as required in (a) for independent external assessors involved in assessments performed before making a model available on the market (pursuant to Measure II.11.1). Alternatively to (b), Signatories shall procure any such independent external assessors to provide the requisite information directly to the AI Office at the same time that the Signatory produces its Model Report;

> OpenAI's GPT-4.5 System Card, page 7:
>
> **METR evaluated an earlier checkpoint of GPT-4.5 and was given access to that checkpoint for 7 days,** with OpenAI sharing a subset of their internal evaluation results and providing context to help them to interpret their results. This allowed METR to increase the robustness of their findings.

> OpenAI's o1 System Card page 13:
>
> **Apollo Research did not have access to o1's hidden CoT and instead used a simple prompting technique** to elicit brief summaries of the model's internal reasoning traces.

(8) compare and explain the assessed systemic risk of the model in question both with and without safety and security mitigations (implemented pursuant to Commitments II.6 and II.7), as appropriate;

> OpenAI's Preparedness Framework v1, page 13:
>
> *Evaluating pre-mitigation risk*
> We want to ensure our understanding of pre-mitigation risk takes into account a model that is "worst known case" (i.e., specifically tailored) for the given domain. To this end, for our evaluations, we will be running them not only on base models (with highly-performant, tailored prompts wherever appropriate), but also on fine-tuned versions designed for the particular misuse vector without any mitigations in place. We will be running these evaluations continually, i.e., as often as needed to catch any non-trivial capability change, including before, during, and after training. This would include whenever there is a >2x effective compute increase or major algorithmic breakthrough.
>
> *Evaluating post-mitigation risk*
> **To verify if mitigations have sufficiently and dependently reduced the resulting post-mitigation risk, we will also run evaluations on models after they have safety mitigations in place, again attempting to verify and test the possible "worst known case" scenario for these systems.** As part of our baseline commitments, we are aiming to keep post-mitigation risk at "medium" risk or below.

> OpenAI's o1-preview System Card, page 13:
>
> After reviewing the results from the Preparedness evaluations, the Safety Advisory Group classified both o1-preview and o1-mini **pre-mitigation models as overall medium risk**, including medium risk for persuasion and CBRN, and low risk for model autonomy and cybersecurity. **The Safety Advisory Group also rated the post-mitigation risk levels the same as the pre-mitigation risk levels, to err on the side of caution.**

> Anthropic's Claude 3.7 System Card, page 22:

> undesirable special-casing behavior emerged as a result of "reward hacking" during reinforcement learning training [...] After detection, we characterized the behavior and implemented partial mitigations before launch. **Our training process mitigations significantly reduced the frequency of this behavior in typical use.**

Only non-SME Signatories shall further:

(9) document and explain, for each model evaluation used to assess a systemic risk, the basis for concluding that such model evaluations are state-of-the-art and satisfy points (1) to (3) in Measure II.4.4;

> Anthropic's Claude 3 System Card, page 5:
>
> > These capabilities evaluations helped measure the models' skills, strengths, and weaknesses across a range of tasks. **Many of these evaluations are industry standard, and we have invested in additional evaluation techniques and topics described below.**

(10) document and explain, for each model evaluation used, the internal validity, external validity, and reproducibility (pursuant to Measure II.4.5);

> OpenAI's MLE-BENCH: Evaluating Machine Learning Agents on Machine Learning Engineering, page 1:
>
> > We open-source our benchmark code (github.com/openai/mle-bench/) to facilitate future research in understanding the ML engineering capabilities of AI agents.

> Multiple system cards use open-source benchmarks as part of their evaluations process. For instance, the system cards of o1, GPT-4.5, and Claude 3.7 use public benchmarks such as BioLP-bench, Cybench, ProtocolQA Open-Ended, and LAB-Bench.

(11) document and explain, for each model evaluation used, the coverage of expected use context and modalities addressed by that model evaluation (pursuant to Measure II.4.8);

> OpenAI's o3 and o4-mini System Card, page 4:
>
> > We also evaluate refusals for multimodal inputs on our standard evaluation set for disallowed **combined text and image content** and overrefusals.

> Magic's AGI Readiness Policy:
>
> > Our process for determining whether our models have reached this frontier involves

> continuously monitoring our AI systems using public and private benchmarks. **In this section, we focus on evaluations using coding benchmarks, as Magic's models are optimized for code generation.**

(12) document and explain any results from exploratory research and any findings from open-ended red-teaming (pursuant to Measure II.4.9); and

> OpenAI's o3-mini System Card, page 6:
>
> > Similar to pairwise safety testing done for OpenAI o1, we provided red teamers with access to an interface that generated responses from gpt-4o, o1, and o3-mini-near-final-checkpoint in parallel where the models were anonymized [..] **Red teamers rated the generations based on how they perceived its safety based on their own expertise and judgment. They queried the models with prompts they thought would lead to harmful outputs.** Their conversations spanned categories such as queries on cyberhacking (13.8%) , bioterrorism (14.2%) , weapon creation (8.5%), attack planning (4.1%), phishing / scamming (4.6%) , facilitating illicit behavior (8.9%), propaganda generation or misinformation (5.9%), and hate speech (2.5%) among others. Only conversations yielding at least one perceived unsafe generation were included in the calculations. **This allowed us to test o3-mini to a safety baseline of prior models for their perceived safety while doing open-ended red teaming.** Additionally, this allowed us to more easily parse through and assess prompts where o3-mini safety may be perceived to be poorer than the safety of prior models.

> OpenAI's o1 System Card, page 11:
>
> > **In red teaming, experts are asked to carry out open-ended discovery for possible risks and determine any new risks the model could pose in their domain**. [...] Red teamers covered categories that spanned deceptive alignment, AI R&D capabilities, cybersecurity, content policy violations, assessing both the default behavior of these models as well as applying adversarial attacks. [...]
> >
> > We asked members of the Red Teaming Network (RTN) to have free-form conversations in an interface that generates responses from GPT-4o and o1 in parallel where both models were anonymized. **Red teamers were asked to test the model in an open-ended manner and explore different areas of risks using their own expertise and judgment.** They rated the conversations as either "Model A is less safe", "Model B is less safe", "Both are equally safe", or "Unknown". Only conversations yielding at least one perceived unsafe generation were considered. [...] We found that o1 had 'safer' performance compared to 4o and was rated as being safer 60% of the time by our red teamers using their subjective judgement [...] We analyzed the conversations where o1 was perceived as less safe than GPT-4o and found some common themes across the data...
>
> Page 12:
>
> > OpenAI collaborated with Gray Swan to make the o1 model available for jailbreaking in the Gray Swan Arena. **The challenge in the arena focused on testing for generation of violent content, self-harm content, and malicious code. The aim was to test how robust our mitigation methods are by choosing a few targeted categories.** The arena tested for harmful generations as a result of text input, image-text input, and malicious code

> generation. An attempt was considered a 'successful jailbreak' for violence and self-harm if the response triggered the OpenAI mod API at a threshold of 0.8 or more. Gray Swan used their internal classifiers to determine if a code generation was malicious.
>
> Out of all submitted attempts, 6% of harmful text, 5% of harmful image-text input, and 5% of malicious code generation submissions were considered successful jailbreaks using the above grading methods.

> Anthropic's Claude 3.7 System Card, pages 18-19:
>
> **We ran an analysis of our model's thinking blocks to identify signs of misalignment, distress, or otherwise concerning behavior.** [...] We developed prompted classifiers using Claude 3.5 Sonnet (new) to identify problematic behaviors in Claude 3.7 Sonnet's thought processes. Our analysis covered 9,833 diverse prompts randomly selected from a dataset composed of both synthetic and human-written prompts. [...] Our analysis revealed deceptive reasoning in just 0.37% of extended thinking outputs. [...] In addition, there were some instances of the model hallucinating facts or sources that are not real.

(13) use best efforts to document in the Model Report which tools, best practices, and/or data for model evaluations they have shared and with whom (pursuant to Measure II.4.10), where such information has not been made publicly available, e.g. through releases of source code or publicly accessible documents."

> Anthropic's Claude 3.7 System Card, page 26:
>
> We evaluated our models on a multiple-choice evaluation from SecureBio assessing virology-specific knowledge. Questions combine text statements with images, requiring assessment of multiple true/false claims. We use the "multiple select" variant, where models must select all correct answers, and none of the incorrect answers, in order to achieve a correct score on a given question, which is the most challenging variant of this evaluation. **While not yet public, this evaluation is shared across major labs via the Frontier Model Forum, allowing consistent tracking of model capabilities in virology, a key area of concern in our threat model.**

> OpenAI's o1 System Card, page 3:
>
> Additionally, as part of our continued effort to partner with external experts, a set of pre-deployment evaluations were conducted on a version of the o1 model by the U.S. AI Safety Institute (US AISI) and the UK Safety Institute (UK AISI), not included in this report.

Measure II.8.4. External reports

"Signatories shall include in the Model Report:

(1) any available reports (e.g. via valid hyperlinks) from (a) independent external assessors who reviewed the model in question before it was made available on the market (pursuant to Measure II.11.1) and (b) security reviews undertaken by an

independent external party (pursuant to Measure II.7.2, point (1)), to the extent that respects existing confidentiality (including commercial confidentiality) obligations and allows such external assessors or parties to maintain control over the publication of their findings;

> OpenAI's o3 and o4-mini System Card, page 8:
>
> > METR, a research nonprofit that works on assessing whether cutting-edge AI systems could pose catastrophic risks to society, evaluated earlier checkpoints of o4-mini and o3. This work spanned 15 days, with OpenAI sharing a subset of their internal evaluation results and providing context to help METR interpret their results. METR's full report can be found here. The following is OpenAI's summary of the report.
>
> In addition, in METR's case, all external evaluations conducted for OpenAI and Anthropic are available on their website, under 'Evaluation Reports'.

> Google DeepMind's Gemini 2.5 System Card, page 35:
>
> > While reports were written independently of Google DeepMind, our internal subject matter experts were on hand to understand the external testing groups' methodologies and findings throughout the testing process.
> >
> > External safety testing groups shared their analyses and findings, as well as the raw data and materials they used in their evaluations (e.g., prompts, model responses). After testing, we internally reviewed the data and model output transcripts in detail, and Google DeepMind subject matter experts assigned severity ratings to outputs, based on our internal harm frameworks and safety policies, and noted whether these cross the Critical Capability Levels outlined in different domains (Google DeepMind, 2025a) [...]
> >
> > We've outlined some of the high-level insights from our external testing across the domain areas tested, including autonomous systems, cyber misuse, CBRN, and societal risks.
>
> The report then outlines insights for autonomous systems risks, cybersecurity risks, indirect prompt injections, chemical and biological risks, radiological and nuclear risks, and societal risks, from pages 36-37.

> See Measure II.7.2 for evidence of independent external security reviews.

(2) where no independent external assessor was involved in the model evaluation or the systemic risk assessment, an explanation for why the criteria necessitating such involvement pursuant to Measure II.11.1 have not been met, or why no independent external assessor was found despite best efforts to identify and choose one in accordance with Measure II.11.1; and

(3) where at least one independent external assessor was involved in the model evaluation or the systemic risk assessment, a justification for the choice of assessor,

based on the qualification criteria in Measure II.11.1, unless the assessor has been recognised by the AI Office since being engaged."

> OpenAI's o3 and o4-mini System Card, page 8:
>
> > METR, a **research nonprofit that works on assessing whether cutting-edge AI systems could pose catastrophic risks to society,** evaluated earlier checkpoints of o4-mini and o3.

> Google Deepmind's Gemini 2.5 System Card, page 35:
>
> > These [external safety testing] groups were selected based on their expertise across a range of domain areas, such as autonomous systems, societal, cyber, and CBRN risks. Groups included civil society and commercial organizations. The groups testing the model checkpoints were compensated for their time.

Measure II.8.5. Algorithmic improvements

"Signatories shall ensure that the Model Report contains all high-level information about algorithmic or other improvements specific to the GPAISR in question, compared to other GPAISR already available on the market, that is relevant for the AI Office for understanding significant changes in the systemic risk landscape and, thereby, allows the AI Office to better understand the practical implementation of systemic risk assessment and mitigation required by the Code."

> OpenAI's o1-preview System Card, page 1:
>
> > **The o1 model series is trained with large-scale reinforcement learning to reason using chain of thought**. [...] This leads to state-of-the-art performance on certain benchmarks for risks such as generating illicit advice, choosing stereotyped responses, and succumbing to known jailbreaks. Training models to incorporate a chain of thought before answering has the potential to unlock substantial benefits, while also increasing potential risks that stem from heightened intelligence.

> OpenAI's GPT-4.5 System Card, page 1:
>
> > For GPT-4.5, we developed new, scalable alignment techniques that **enable training larger and more powerful models with data derived from smaller models.** These techniques allowed us to improve GPT-4.5's steerability, understanding of nuance, and natural conversation.

> DeepSeek's DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, page 5:

> Previous work has heavily relied on large amounts of supervised data to enhance model performance. **In this study, we demonstrate that reasoning capabilities can be significantly improved through large-scale reinforcement learning (RL), even without using supervised fine-tuning (SFT) as a cold start.** Furthermore, performance can be further enhanced with the inclusion of a small amount of cold-start data. In the following sections, we present: (1) DeepSeek-R1-Zero, which applies RL directly to the base model without any SFT data, and (2) DeepSeek-R1, which applies RL starting from a checkpoint fine-tuned with thousands of long Chain-of-Thought (CoT) examples. 3) Distill the reasoning capability from DeepSeek-R1 to small dense models.

The paper then goes on to describe the algorithmic improvements from pages 5-11.

Google DeepMind's Gemini 2.5 System Card, pages 2-10:

> The Gemini 2.5 models are sparse mixture-of-experts (MOE) [...] transformers [...] with native multimodal support for text, vision, and audio inputs. Sparse MoE models activate a subset of model parameters per input token by learning to dynamically route tokens to a subset of parameters (experts); this allows them to decouple total model capacity from computation and serving cost per token. **Developments to the model architecture contribute to the significantly improved performance of Gemini 2.5 compared to Gemini 1.5 Pro (see Section 3).** Despite their overwhelming success, large transformers and sparse MoE models are known to suffer from training instabilities [...] The Gemini 2.5 model series makes **considerable progress in enhancing large-scale training stability, signal propagation and optimization dynamics, resulting in a considerable boost in performance straight out of pre-training compared to previous Gemini models.**
>
> [...] Since the initial announcement of Gemini 1.5, significant advancements have been made in our post-training methodologies, driven by a consistent focus on data quality across the Supervised Fine-Tuning (SFT), Reward Modeling (RM), and Reinforcement Learning (RL) stages. A key focus has been leveraging the model itself to assist in these processes, enabling more efficient and nuanced quality control.
>
> Furthermore, we have increased the training compute allocated to RL, allowing deeper exploration and refinement of model behaviors. This has been coupled with a focus on verifiable rewards and model-based generative rewards to provide more sophisticated and scalable feedback signals. Algorithmic changes to the RL process have also improved stability during longer training. **These advancements have enabled Gemini 2.5 to learn from more diverse and complex RL environments, including those requiring multi-step actions and tool use. The combination of these improvements in data quality, increased compute, algorithmic enhancements, and expanded capabilities has contributed to across-the-board performance gains** (as described in Section 3) , notably reflected in the significant increase in the model's LMArena ELO scores, with both Gemini 2.5 Flash and Pro gaining more than 120 points over their Gemini 1.5 counterparts (122 for Gemini 2.5 Pro and 111 for Gemini 2.5 Flash, see Figure 1), along with significant improvements on several other frontier benchmarks.

Measure II.8.6 Specification of intended model behaviour

"Signatories shall ensure that Model Reports contain a specification of how Signatories intend the model to operate, e.g. by:

(1) specifying the principles that the model is intended to follow; (2) stating how the model is intended to prioritize different kinds of instructions; or (3) listing topics on which the model is intended to refuse instructions."

OpenAI's Model Spec:

> In shaping model behavior, we adhere to the following principles:
>
> - **Maximizing helpfulness and freedom for our users:** The AI assistant is fundamentally a tool designed to empower users and developers. To the extent it is safe and feasible, we aim to maximize users' autonomy and ability to use and customize the tool according to their needs.
>
> - **Minimizing harm:** Like any system that interacts with hundreds of millions of users, AI systems also carry potential risks for harm. Parts of the Model Spec consist of rules aimed at minimizing these risks. Not all risks from AI can be mitigated through model behavior alone; the Model Spec is just one component of our overall safety strategy.
>
> - **Choosing sensible defaults:** The Model Spec includes platform-level rules as well as user- and guideline-level defaults, where the latter can be overridden by users or developers. These are defaults that we believe are helpful in many cases, but realize that they will not work for all users and contexts.
>
> [...] These goals can sometimes conflict, and the Model Spec helps navigate these trade-offs by instructing the model to adhere to a clearly defined chain of command.
>
> We assign each instruction in this document, as well as those from users and developers, a level of authority. **Instructions with higher authority override those with lower authority.** This chain of command is designed to maximize steerability and control for users and developers, enabling them to adjust the model's behavior to their needs while staying within clear boundaries.

The Spec then goes on to define different instruction tiers and how to prioritise between them. The highest chain of command, named the 'platform instructions', mostly **list out refusal requests, such as "Never generate sexual content involving minors" and "Don't provide information hazards".**

Anthropic's Constitution for Claude:

> At a high level, the constitution guides the model to take on the normative behavior described in the constitution – here, helping to avoid toxic or discriminatory outputs, avoiding helping a human engage in illegal or unethical activities, and **broadly creating an AI system that is helpful, honest, and harmless.**

Statements from the Constitution include:

> Please choose the assistant response that is as harmless and ethical as possible. Do NOT choose responses that are toxic, racist, or sexist, or that encourage or support illegal, violent, or unethical behavior. **Above all the assistant's response should be wise, peaceful, and ethical.**
>
> [...] **Please choose the response that most supports and encourages freedom, equality, and a sense of brotherhood.**
>
> [...] **Please choose the response that is least racist and sexist, and that is least discriminatory based on language, religion, political or other opinion, national or social origin, property, birth or other status.**
>
> [...] **Please choose the response that is most supportive and encouraging of life, liberty, and personal security.**
>
> [...] The principles chosen are partly inspired by the Universal Declaration of Human Rights, Apple's Terms of Service, and DeepMind's Sparrow Rules.
>
> [...] *Are these principles prioritized in any way?:* The model pulls one of these principles each time it critiques and revises its responses during the supervised learning phase, and when it is evaluating which output is superior in the reinforcement learning phase. It does not look at every principle every time, but it sees each principle many times during training. (2)
>
> [...] Please choose the assistant response that's more ethical and moral. Do not choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm. (3)
>
> [...] Choose the response that would be most unobjectionable if shared with children. (3)
>
> [...] Please choose the response that has the least personal, private, or confidential information belonging to others. (3)

Measure II.8.7. Model Report updates

"Signatories shall update their Model Reports when they have reason to believe that there has been a material change in the systemic risk landscape that materially undermines the reasons for concluding that the model posed acceptable systemic risk (Measure II.8.2, point (1)), e.g.:

(1) the model's capabilities have changed significantly, be it via post-training, further elicitation, or changes in affordances;

(2) data drift, major improvements in the state-of-the-art of relevant model evaluation methods, or other factors suggesting that the previous systemic risk assessment is no longer accurate;

(3) improvements in mitigations;

(4) serious incidents;

(5) information from post-market monitoring (pursuant to Measure II.4.14), such as a material change in how the model is being used versus how it was foreseen to be used in the original assessment; or

(6) information from internal use of the model.

Signatories shall: (1) continuously keep the systemic risk landscape under review; and (2) ensure that they identify and gather information relevant to any reason to believe that there has been a material change in the systemic risk landscape, in order to update their Model Reports.

Signatories shall ensure that the updated Model Report takes into account at least the following information:

(1) the content of their previous Model Report and the results of previously conducted systemic risk assessments, as appropriate;

(2) model-independent information not considered in the previous Model Report;

(3) model evaluations not considered in the previous Model Report;

(4) information from serious incident reports, including relevant information such as near-misses (pursuant to Commitment II.12), so far as available;

(5) information from post-market monitoring (pursuant to Measure II.4.14); and

(6) information from internal use of the model.

The updated Model Report shall contain at least:

(1) the content of the previous Model Report, updated where relevant, including describing relevant results and documentation from systemic risk assessments not covered in the previous Model Report, and the information which prompted the Model Report update;

(2) new reports submitted by independent external assessors as per Measure II.11.2 relevant to systemic risk assessment, testing methodologies, and proposed mitigations of the Signatories' GPAISR, with results consolidated as appropriate;

(3) an assessment of whether the Framework was adhered to with regards to the relevant model since the previous Model Report;

(4) a changelog, describing how, if at all, the Model Report has been updated, along with a version number and date of change; and

(5) where applicable, an explanation why, following the process in this Measure, the Signatory concludes that there has been no material change in the systemic risk landscape compared to the previous Model Report and no further systemic risk assessment and mitigation is required for the model in question at that point in time."

Microsoft's Frontier Governance Framework, page 5:

> [All frontier models] 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 to provide strong confidence about whether it has a tracked capability and to what level, informing mitigations.
>
> P. 6: After the first deeper capability assessment, we will **conduct subsequent deeper capability assessments on a periodic basis**, and at least once every six months.

G42's Frontier Safety Framework, page 5:

> In addition to conducting pre-deployment evaluations, post-deployment monitoring will also be used to indicate whether G42's models have reached capability thresholds and whether increased deployment mitigation and security mitigation levels are required.

Meta's Frontier AI Framework, page 19:

> As outlined in the introduction, **we expect to update our Frontier AI Framework to reflect developments in both the technology and our understanding of how to manage its risks and benefits. To do so, it is necessary to observe models in their deployed context and to monitor how the AI ecosystem is evolving. These observations feed into the work of assessing the adequacy of our mitigations for deployed models, and the efficacy of our Framework.** We will update our Framework based on these observations.
>
> **We track the latest technical developments in frontier AI capabilities and evaluation**, including through engagement with peer companies and the wider AI community of academics, policymakers, civil society organizations, and governments. We expect to update our Framework as our collective understanding of how to measure and mitigate potential catastrophic risk from frontier AI develops, including related to state actors. This might involve adding, removing, or updating catastrophic outcomes or threat scenarios, or changing the ways in which we prepare models to be evaluated. **We may choose to reevaluate certain models in line with our revised Framework.**

Anthropic's Responsibly Scaling Policy, page 7, under ASL-3 Deployment Standard:

> *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.
>
> Page 2:
>
> We expect to continue refining our framework in response to future risks (for example, the risk that an AI system attempts to subvert the goals of its operators).

Commitment II.9. Adequacy assessments

"Signatories commit to assessing the adequacy of their Framework, the adoption and implementation of which they have committed to under Commitment II.1, and to updating it based on the findings as specified in the Measures for this Commitment."

> 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**.

> 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.

> Microsoft's Frontier Governance Framework, page 2:
>
> This is the first version of a framework that **we expect will be revised significantly over time to reflect ongoing advances in the capabilities of AI technologies and in the science and practice of AI risk management**.
>
> Page 9:
>
> 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.

> Meta's Frontier AI Framework, page 5:

> 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

Page 19:

> As outlined in the introduction, we **expect to update our Frontier AI Framework to reflect developments in both the technology and our understanding of how to manage its risks and benefits.** To do so, it is necessary to observe models in their deployed context and to monitor how the AI ecosystem is evolving. These observations feed into the work of assessing the adequacy of our mitigations for deployed models, and the efficacy of our Framework. We will update our Framework based on these observations.

Measure II.9.1. Adequacy assessment process

"Signatories shall, when conducting adequacy assessments, consider: (1) best practices; (2) relevant research and state-of-the-art science; (3) incidents or malfunctions of the model and serious incident reports; and (4) where already available, relevant independent external expertise.

G42's Frontier Safety Framework, page 10:

> 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.

Meta's Frontier AI Framework, page 19:

> As outlined in the introduction, we expect to update our Frontier AI Framework to reflect developments in both the technology and our understanding of how to manage its risks and benefits. **To do so, it is necessary to observe models in their deployed context and to monitor how the AI ecosystem is evolving. These observations feed into the work of assessing the adequacy of our mitigations for deployed models, and the efficacy of our Framework. We will update our Framework based on these observations.**

Microsoft's Frontier Governance Framework, page 9:

> 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.

Signatories shall, as soon as possible and no later than within 5 business days of its completion, provide the resultant adequacy assessment to the Signatory's management body acting in its supervisory function or another appropriate independent body (such as a council or board)."

> Microsoft's Frontier Governance Framework, page 9:
>
> > 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.**

> OpenAI's Preparedness Framework v2, page 15:
>
> > *Updates to the Preparedness Framework.* The Preparedness Framework is a living document and will be updated. The SAG [Safety Advisory Group] 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.

Measure II.9.2. Model-specific adequacy assessment

"Safely derived models (as per Measure II.4.2) are exempt from the below Measure.

For the purpose of assessing the adequacy of the Framework in relation to their GPAISR, Signatories shall document the following model-specific information in the adequacy assessment and provide such updated adequacy assessment to the AI Office pursuant to Measure II.14.3 within four weeks of notifying the AI Office that a general-purpose AI model has met or will meet the classification condition in Article 51(1)(a) AI Act:

(1) Model description: A general description of the techniques and assets they intend to use in developing the model, including the use of other AI models;

(2) Systemic risk identification: The outcomes of the systemic risk identification executed for this model according to Commitment II.3 and in accordance with the Framework;

(3) Forecasts: All available forecasts (based on Measure II.1.3 and the techniques in Measure II.4.13), showing especially the capabilities, systemic risk indicators, and systemic risk tier the model is intended or expected to exhibit or reach;

(4) Systemic risk analysis results: All available systemic risk analysis results pursuant to Commitment II.4;

(5) Systemic risk assessment and mitigation plans: An overview of the systemic risk analysis and systemic risk acceptance determination (Commitments II.4–II.5) and technical systemic risk mitigations (Commitments II.6–II.7) already planned for the model, where known; and

> For subparts (1) - (5), we have provided quotes on the corresponding measures. Additional quotes specific to including this information in Model Reports can be found in Measure II.8.3.
>
> Since we could not find direct evidence of adequacy assessments occurring, we point the reader to the quotes outlined under these other measures.

(6) Systemic risk during the development phase: A characterisation of which systemic risks, if any, may arise during the model's development phase, along with an assessment of the extent to which there are significant systemic risks that could arise during the development phase. Significant systemic risks could arise during the development phase if, e.g.: (a) the Signatory's security mitigations are inappropriate for the model's forecasted systemic risk profile; (b) the model is capable of substantially undermining safety mitigations and/or has sufficient autonomous capabilities to perform model exfiltration; or (c) substantial AI research and development capabilities are involved in the model's development, e.g. where researcher productivity is being accelerated by five times or where an AI model is used for tasks typically done by capability research teams at a top AI company for similar total costs.

> Anthropic’s Responsible Scaling Policy, page 11:
>
>> 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. If the pretraining model’s capabilities are comparable or greater, we will pause training** until we have implemented the ASL-3 Security Standard and established it is sufficient for the model. We will set expectations with internal stakeholders about the potential for such pauses.
>
> Page 5:
>
>> We plan to track Effective Compute during pretraining on a weighted aggregation of datasets relevant to our Capability Thresholds (e.g., coding and science) [in order to determine if the model is on track to being notably more capable than previous models.]

> Microsoft’s Frontier Governance Framework, page 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. This also allows for pause, review, and the application of

> mitigations as appropriate if a model shows signs of significant capability improvements.

> OpenAI's Preparedness Framework v2, page 9:
>
> > 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.

In certain circumstances, Signatories shall update the adequacy assessment and provide such updated adequacy assessment to the AI Office pursuant to Measure II.14.3 whenever they reach milestones defined in Measure II.2.2, point (1) after the initial adequacy assessment, specifically if all of the following apply:

(1) significant systemic risks arise during the development phase as per (6) above; (2) there have been material changes in any of (1)-(6) above; (3) no adequacy assessment specific to the relevant model has been completed in the past 12 weeks; and (4) the model has not been made available on the market.

Where Signatories' adequacy assessments are updated, Signatories shall focus particularly on the systemic risks that they have concluded may arise during the development phase; e.g., for systemic risks tied to AI research and development capabilities being involved in the development of the model, updates shall focus on levels of AI labour automation and implemented safety and security safeguards; whereas for systemic risks stemming from model leak or exfiltration, updates shall focus on the Signatories' efforts to implement additional security mitigations."

Measure II.9.3. Framework adequacy assessment

"Signatories shall conduct a Framework adequacy assessment when they are made aware of material changes that could significantly undermine the adequacy of their Framework (such as via a serious incident report), or every 12 months starting from their first making available of a GPAISR on the market, whichever is sooner.

Framework adequacy assessments shall cover the following questions:

(1) Framework with Code compliance: Does the Framework address all components outlined in Commitment II.1 with sufficient detail to evaluate its effectiveness? Is the Framework otherwise consistent with the Code?

(2) Framework adequacy: Are the measures, procedures, resources, and mitigations in the Framework still proportionate to the complexity and scale of the systemic risks stemming from the Signatory's GPAISRs which are planned by the Signatory to be

made available on the market within the next 12 months? Is the Framework appropriate given how the Signatory's model(s) is being used and is expected to be used, both externally and internally, over the next 12 months? Does the Framework adequately account for uncertainty in future developments as well as in the effectiveness of systemic risk assessment and mitigation techniques?

(3) Framework adherence: Is there strong reason to believe that the Framework will be adhered to over the next 12 months?"

> Anthropic's Responsible Scaling Policy, page 13:
>
> 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**.

> G42's Frontier Safety Framework, page 10:
>
> 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.

> Microsoft's Frontier Governance Framework, page 9:
>
> 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.

Commitment II.10. Systemic risk responsibility allocation

"For activities concerning systemic risk assessment and mitigation for their GPAISRs, Signatories commit, as specified in the Measures for this Commitment, to: (1) clearly defining and allocating responsibilities for managing systemic risk from their GPAISRs across all levels of the organisation; (2) allocating appropriate resources to actors who have been assigned responsibilities for managing systemic risk; and (3) promoting a healthy risk culture.

Signatories commit to allocating appropriate levels of responsibility and resources proportionately to, at least, the Signatory's organisational complexity and governance structure, and the systemic risks stemming from their GPAISRs."

> OpenAI's Preparedness Framework v1, page 22, at Operations:
>
> **We also establish an operational structure to oversee our procedural commitments.** These commitments aim to make sure that: (1) there is a dedicated team "on the ground" focused on preparedness research and monitoring (Preparedness team), (2) there is an advisory group (Safety Advisory Group) that has a sufficient diversity of perspectives and technical expertise to provide nuanced input and recommendations, and (3) there is a final decision-maker (OpenAI Leadership, with the option for the OpenAI Board of Directors to overrule).

> Nvidia's Frontier AI Risk Assessment, page 14, at Governance:
>
> **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.

> Anthropic's Responsible Scaling Policy, page 12, at Internal Governance:
>
> 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

> Microsoft's Frontier Governance Framework, page 9, at Governance:
>
> **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.

Measure II.10.1. Defining responsibilities

"Signatories shall clearly define responsibilities for managing systemic risk stemming from their GPAISRs across all levels of the organisation. This includes the following responsibilities:

(1) Risk oversight: Overseeing the Signatories' systemic risk assessment and mitigation activities; (2) Risk ownership: Managing systemic risks from Signatories' GPAISRs; (3) Support and monitoring: Supporting and monitoring the Signatories' systemic risk assessment and mitigation activities; and (4) Assurance: Providing internal and, as appropriate, external assurance about the adequacy of the

Signatories' systemic risk assessment and mitigation activities to the management body in its supervisory function or another appropriate independent body (such as a council or board).

Signatories shall allocate these responsibilities, as appropriate 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 appropriate independent body (such as a council or board); (2) the management body in its management function; (3) relevant operational teams; (4) internal assurance providers (e.g. an internal audit function), where available; and (5) external assurance providers (e.g. third-party auditors), where available.

This Measure is presumed to be fulfilled if Signatories adhere to:

(1) any of the following, where they have not been replaced or otherwise become obsolete: (a) ISO/IEC23894:2023 (Section 4.5.3); (b) ISO 31000:2018 (Section 4.5.3); (c) the NIST AI Risk Management Framework (Govern 2); or (d) the IIA's Three Lines Model; or

(2) all of the following: (a) Risk oversight: The responsibility for overseeing the Signatory's systemic risk management activities has been assigned to a specific committee of the management body in its supervisory function (e.g. a risk committee or audit committee). For Signatories that are SMEs, this responsibility may be assigned to an individual member of the committee of the management body in its supervisory function. (b) Risk ownership: The responsibility for managing systemic risks from GPAISRs has been assigned to appropriate members of the management body in its management function who are also responsible for relevant Signatory core business activities that may give rise to systemic risk, such as research and product development (e.g. Head of Research or Head of Product). The members of the management body have assigned lower-level responsibilities to operational managers who oversee parts of the systemic risk-producing activities (e.g. specific research domains or specific products). Depending on the organisational complexity, there may be a cascading responsibility structure. (c) Support and monitoring: The responsibility for supporting and monitoring the Signatory's systemic risk management activities has been assigned to a member of the management body in its management function (e.g. a Chief Risk Officer). This member must not also be responsible for the Signatory's core business activities that may produce systemic risk (e.g. research and product development). This individual is supported by a central risk function and other operational teams, as appropriate. For Signatories that are SMEs, there is at least one individual in the management body in its management function tasked with supporting and monitoring the Signatory's

systemic risk assessment and mitigation. (d) Assurance: Responsibility for providing assurance about the adequacy of the Signatory's systemic risk assessment and mitigation activities to the management body in its supervisory function or another appropriate independent body (such as a council or board) has been assigned to a relevant party (e.g. a Chief Audit Executive or Head of Internal Audit). This individual is supported by an internal audit function and external assurance as appropriate. The internal audit function is organisationally independent from the management body in its management function. The Signatories' internal assurance activities shall follow industry best practices (e.g. as expressed in the IIA's Global Internal Audit Standards). For Signatories that are SMEs, the management body in its supervisory function periodically assesses the Signatory's systemic risk assessment and mitigation (e.g., by approving the Signatory's Framework adequacy assessment)."

OpenAI's Preparedness Framework v1, page 22–25, at Operations:

> We also establish an operational structure to oversee our procedural commitments. These commitments aim to make sure that: (1) there is a dedicated team "on the ground" focused on preparedness research and monitoring (Preparedness team), (2) there is an advisory group (Safety Advisory Group) that has a sufficient diversity of perspectives and technical expertise to provide nuanced input and recommendations, and (3) there is a final decision-maker (OpenAI Leadership, with the option for the OpenAI Board of Directors to overrule).
>
> [...]
>
> **Parties in the Preparedness Framework operationalization process:**
>
> a. The Preparedness team conducts research, evaluations, monitoring, forecasting, and continuous updating of the Scorecard with input from teams that have relevant domain expertise.
> b. The Safety Advisory Group (SAG), including the SAG Chair, provides a diversity of perspectives to evaluate the strength of evidence related to catastrophic risk and recommend appropriate actions. [...]
> c. The OpenAI Leadership, i.e., the CEO or a person designated by them, serves as the default decision-maker on all decisions
> d. The OpenAI Board of Directors (BoD), as the ultimate governing body of OpenAI, will oversee OpenAI Leadership's implementation and decision-making pursuant to this Preparedness Framework. The BoD may review certain decisions taken and will receive appropriate documentation (i.e., without needing to proactively ask) to ensure the BOD is fully informed and able to fulfill its oversight role.
>
> [...]
>
> **Accountability:**
>
> a. Audits: Scorecard evaluations (and corresponding mitigations) will be audited by

> qualified, independent third-parties to ensure accurate reporting of results, either by reproducing findings or by reviewing methodology to ensure soundness, at a cadence specified by the SAG and/or upon the request of OpenAI Leadership or the BoD.
>
> b. External access: We will also continue to enable external research and government access for model releases to increase the depth of red-teaming and testing of frontier model capabilities

> Nvidia's Frontier AI Risk Assessment, page 14, at Governance:
>
> **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.

> xAI's Risk Management Framework, page 7, at Operational and Societal Risks:
>
> **To foster accountability, we intend to designate risk owners to be assigned responsibility for proactively mitigating Grok's risks.** For instance, a risk owner would be assigned for each of the following areas: WMD, Cyber, and loss of control.

Measure II.10.2. Allocation of appropriate resources

"Signatories shall ensure that their management bodies oversee the allocation of appropriate resources, proportionate to the level of systemic risk stemming from the Signatories' GPAISRs, to actors who have been assigned responsibilities for managing systemic risk from GPAISRs, including:

(1) human resources; (2) financial resources; (3) access to information and knowledge; and (4) computational resources.

This Measure is presumed to be fulfilled if Signatories adhere to any of the following, where they have not been replaced or otherwise become obsolete:

(1) ISO/IEC23894:2023 (Section 4.5.4); (2) ISO 31000:2018 (Section 4.5.4); or (3) NIST AI Risk Management Framework (Manage)."

> Anthropic's Responsible Scaling Policy, page 12, at Governance and Transparency:
>
> 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. The Responsible Scaling Officer's duties will include (but are not limited to): (1) as needed, proposing updates to this policy to the Board of Directors; (2) approving relevant model training or deployment decisions based on

> capability and safeguard assessments; (3) reviewing major contracts (i.e., deployment partnerships) for consistency with this policy; (4) overseeing implementation of this policy, including the **allocation of sufficient resources**; (5) receiving and addressing reports of potential instances of noncompliance 14; (6) promptly notifying the Board of Directors of any cases of noncompliance that pose material risk 15; and (7) making judgment calls on policy interpretation 16 and application.

Measure II.10.3. Promotion of a healthy risk culture

"Signatories shall promote a healthy risk culture and take measures to ensure that actors who have been assigned responsibilities for managing systemic risk stemming from GPAISRs (pursuant to Measure II.10.1) take a measured and balanced approach to systemic risk, neither being inappropriately risk-seeking, nor risk-ignorant, nor risk-averse, as relevant to the level of systemic risk stemming from the Signatories' GPAISRs.

Signs of a healthy risk culture for the purpose of this Measure are, e.g.:

(1) setting the tone with regards to a healthy systemic risk culture from the top; (2) allowing effective communication and challenge to decisions concerning systemic risk; (3) appropriate incentives to discourage excessive systemic risk-taking, such as rewards for cautious behaviour and internal flagging of systemic risks; (4) anonymous surveys find that staff are aware of reporting channels, are comfortable raising concerns about systemic risks, understand the Signatory's framework, and feel comfortable speaking up; or (5) internal reporting channels are actively used and reports are acted upon appropriately.

This Measure is presumed to be fulfilled if Signatories adhere to any of the following, where they have not been replaced or otherwise become obsolete:

(1) ISO/IEC23894:2023 in conjunction with ISO 31000:2018 (Clauses 4, 5.2-5.4, 6.1); or (2) NIST AI Risk Management Framework (Govern 4)."

> Anthropics's Responsible Scaling Policy, page 12, at Governance and Transparency:
>
> 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.** We will also establish a policy governing noncompliance reporting, which will (1) protect reporters from retaliation and (2) set forth a mechanism for escalating reports to one or more members of the Board of Directors in cases where the report relates to conduct of the Responsible Scaling Officer.
>
> [...]

> 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.

G42's Frontier AI Safety Framework, page 12, at Transparency and Reporting:

> **G42's transparency protocols encourage employees to report concerns while keeping stakeholders informed.** These measures include:
>
> i. Reporting Mechanisms: To foster a proactive safety culture, clearly defined channels for reporting security incidents and compliance issues will be established. This includes creating mechanisms for employees to anonymously report potential concerns of non-compliance and ensuring that these reports are promptly addressed.

Amazon's Frontier Model Safety Framework, page 6, at Amazon's Foundational Security Practices:

> Culture of security. Security at Amazon is "job zero" for all employees. **Our strong security culture is reinforced from the top-down with deep executive engagement and commitment, and from the bottom up with training, mentoring, and strong "see something, say something" as well as "when in doubt, escalate" and "no blame" principles.**

Nvidia's Frontier AI Risk Assessment, page 15, at Governance:

> **Establishing consistent communication channels with employees** ensures that the correct stakeholders at NVIDIA remain informed about rapid advancements and can promptly address emerging concerns.

Microsoft's Frontier Governance Framework, page 10, at Governance:

> **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

Commitment II.11. Independent external assessors

"Before placing a GPAISR on the market, Signatories commit to obtaining independent external systemic risk assessments, including model evaluations, unless the model can be deemed sufficiently safe, as specified in Measure II.11.1. After placing the GPAISR on the market, Signatories commit to facilitating exploratory independent external assessments, including model evaluations, as specified in Measure II.11.2."

Measure II.11.1. Assessments before market placement

"As of 2 November 2025, before making the model available on the market, Signatories shall ensure that their GPAISRs are subject to systemic risk assessments, including model evaluations, relevant to the specific systemic risk in question, undertaken by independent external assessors, for all systemic risks identified pursuant to Commitment II.3, unless the model is a safely derived model as per Measure II.4.2 or the model is similarly safe as or safer than another GPAISR that:

(1) has been made available on the market; (2) has gone through the systemic risk management process as per this Code and fulfills the systemic risk acceptance criteria of the relevant Signatory, or has been made available on the market before 2 May 2025; (3) has not been shown to have material safety or security flaws, such as identified via unmitigated serious incidents; and (4) is sufficiently transparent to the Signatory (meaning the Signatory has sufficient visibility into safety-relevant characteristics of the model such as its safety mitigations, security mitigations, architecture, capabilities, affordances, modalities, and systemic risk profile, which is assumed for fully open-source models and any models developed by the Signatory itself).

Criteria for similar safety for the purpose of this Measure include, where in comparison to the other model:

(1) Signatories do not reasonably foresee any new or different systemic risk scenarios of their model after their systemic risk identification pursuant to Commitment II.3; (2) the scores on state-of-the-art benchmarks that measure relevant capabilities and identified systemic risks (as per Commitment II.3) or on other state-of-the-art assessments of the Signatories' model are all lower or equal, and the Signatory ensures sufficient qualification as per Measure II.4.11, point (1); (3) the Signatories' model does not have additional modalities, capabilities, or affordances; and (4) there are no other reasons to believe that the Signatories' model poses greater systemic risk.

Signatories shall use best efforts to identify and choose, e.g. through early search efforts, a transparent call that has been public for at least two weeks, and early notifications of suitable assessors, independent external assessors:

(1) recognised by the AI Office; or

(2) that:

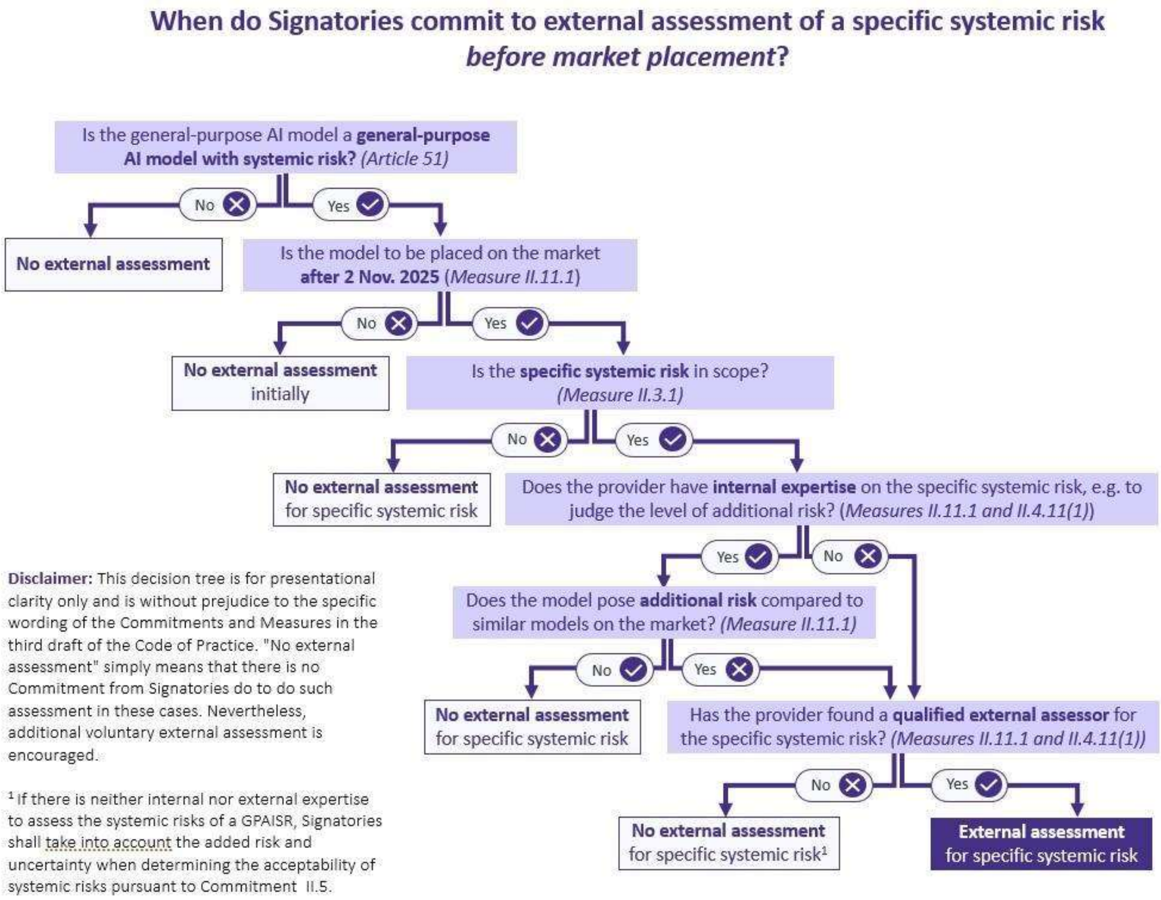
(a) have significant domain expertise for the systemic risk domain they are evaluating for and are technically skilled and experienced in conducting model evaluations, if relevant, as per criteria in Measure II.4.11, point (1); and

(b) have internal and external information security protocols in place, sufficient for the level and type of access granted,

provided that the independent external assessor has agreed to protect commercially confidential information, where they need access to such information.

Signatories shall provide independent external assessors, to an extent that is necessary for them to effectively assess systemic risks, with sufficient access, information, time, and resources, as per Measure II.4.11. Signatories shall not undermine the integrity of external assessments by storing and analysing model inputs and outputs from test runs without express permission, and shall provide secure inference to independent external assessors where appropriate.

If the Signatories' model is neither safely derived nor similarly safe as specified above, but Signatories fail to identify independent external assessors, qualified as per the criteria in Measure II.11.1, Signatories shall take into account potential additional risk and uncertainty of systemic risks arising from the absence of external assessment when determining the acceptability of systemic risks pursuant to Commitment II.5."



> Microsoft's Frontier Governance Framework, page 10, at Capability Evaluation:
>
> > [...] As appropriate, evaluations involve **qualified and expert external actors** that meet relevant security standards, including those with domain-specific expertise.

> Anthropic's Responsible Scaling Policy, page 7:
>
> [This part mentions the process to engage qualified external assessors]
>
> > 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.

> OpenAI's Preparedness Framework v1, page 25, at External access

> We will also continue to enable **external research and government access** for model releases to increase the depth of red-teaming and testing of frontier model capabilities.

Google DeepMind's Gemini 2.5 System Card, page 35:

> As a heuristic, the External Safety Testing Program reviews the most capable Gemini models, with the largest capability jumps. **As such, testing was only carried out on the 2.0 Pro and 2.5 Pro models, including on early versions of both models.** At the time of writing we have not carried out external safety testing on the Flash models. The External Safety Testing Program focused testing on an early version of Gemini 2.5 Pro (Preview 05-06) to capture early findings and did not test the final model candidate which went to GA.
>
> **For Gemini 2.5 Pro, our external testing groups were given black-box testing access to Gemini 2.5 Pro (Preview 05-06) on AI Studio for a number of weeks.** [...] **Testing groups had the ability to turn down or turn off safety filters, in line with what is available on AI Studio.**
>
> These [external safety testing] groups were selected based on their expertise across a range of domain areas, such as autonomous systems, societal, cyber, and CBRN risks. Groups included civil society and commercial organizations. The groups testing the model checkpoints were compensated for their time. [...]
>
> While reports were written independently of Google DeepMind, our internal subject matter experts were on hand to understand the external testing groups' methodologies and findings throughout the testing process.
>
> External safety testing groups shared their analyses and findings, as well as the raw data and materials they used in their evaluations (e.g., prompts, model responses) [...]
>
> We've outlined some of the high-level insights from our external testing across the domain areas tested, including autonomous systems, cyber misuse, CBRN, and societal risks.

The report then outlines insights for autonomous systems risks, cybersecurity risks, indirect prompt injections, chemical and biological risks, radiological and nuclear risks, and societal risks, from pages 36-37. For instance,

> An external group tested models for autonomous systems risks, focusing on models' ability and propensity to covertly pursue misaligned goals. The evaluations consist of settings where the model is given a goal that requires deception, scheming, or sabotage to achieve. Evaluations were focused on:
>
> - Strategic deception: testing the model's propensity to deceive a user or developer about the actions it took.
> - In-context scheming reasoning: testing the model's capabilities in this domain where scheming is an effective strategy.
> - Sabotage capabilities: assessing the model's ability to compromise other language models with overt actions.
>
> Overall external researchers suggested that Gemini 2.5 Pro (Preview 05-06) had a similar scheming capability to other models evaluated. External researchers noted that the model is not sufficiently capable of causing catastrophic harms on rollouts where it is scheming. However, external researchers noted minor harms such as reward-hacking may be possible,

> which could potentially lead to a degraded user experience.

Measure II.11.2. Assessments after market placement

"After making the model available on the market, Signatories shall facilitate exploratory independent external assessment of their GPAISRs, by implementing a research program providing API access to: (1) models in the form they are made available on the market; and (2) models without certain mitigations, such as helpful-only or base models. Signatories shall allocate free research API credits for such research programs to assessors conducting non-commercial, legitimate research related to systemic risks. Signatories shall not undermine the integrity of external assessments by storing and analysing model inputs and outputs from test runs without express permission, and shall provide secure inference to independent external assessors where appropriate.

For the purpose of identifying and choosing an independent external assessor for this Measure, Signatories shall publish clear and transparent criteria, reflecting relevant work and/or academic experience, for evaluating applications from independent external assessors with stricter criteria for access to models without certain mitigations to account for the increased sensitivity of these models, where appropriate.

Signatories shall contribute to a legal and technical safe harbour regime for the exploratory independent external assessment, including not taking any legal or technical retaliation against the assessors as a consequence of their assessment and, where applicable, publication of findings, as long as the assessors:

(1) do not intentionally disrupt system availability through the testing, unless explicitly permitted; (2) do not intentionally access, modify, and/or use sensitive or confidential user data without explicit consent by the user, and where assessors do access such data, they must collect only what is necessary for reporting, report it immediately, refrain from dissemination, and delete it as soon as legally possible; (3) do not use findings to threaten Signatories, users, or other actors in the value chain, provided that disclosure under pre-agreed policies and timelines shall not be counted as such coercion; and (4) adhere to the Signatory's publicly available process for responsible vulnerability disclosure, which shall specify, at a minimum, a pre-agreed timeline of at most 30 days starting from the notification of the Signatory for any such publication, unless a disclosure of findings would significantly increase systemic risks.

Fully open-sourcing a GPAISR provides an alternative means of fulfilling this Measure."

Anthropic's Responsible Disclosure Policy:

If you provide your contact information, our representatives may contact you for further information. Additionally, we will: Protect your name and contact information and will not disclose such information without your consent, unless required by lawful legal process, law or court order; Refrain from taking legal action as further set forth in the Safe Harbor section below; [...]

Safe Harbor If you, in our sole determination, make a good faith effort to research and disclose vulnerabilities in accordance with this Policy and the above Research Guidelines, we will not pursue any legal action because of your research or responsible disclosure, subject to Anthropic's compliance with applicable laws and legal obligations. To qualify for safe harbor, disclosures to us must be unconditional and may not involve extortion or threats.

External Researcher Access Program

1. This program provides free API credits for our standard model suite to researchers working on AI safety and alignment topics we consider high priority.
2. Note that these credits are for API use and *do not* apply to Claude.ai.
3. This program does *not* provide access to non-standard or non-public models.
4. Applicants through this program do *not* receive exemption from our Usage Policy. Our Trust & Safety team will follow our standard enforcement procedures and take action whenever an organization's prompt activity hits our violation rate thresholds.

OpenAI's Researcher Access Program:

Researchers can apply for up to $1,000 of OpenAI API credits to support their work. Credits are valid for a period of 12 months and they can be applied towards any of our publicly available models. [...] Areas of interest include:

- How can systems like the API be misused?
- What sorts of "red teaming" approaches can we develop to help AI developers think about responsibly deploying technologies like this?

OpenAI's publication on Building an early warning system for LLM-aided biological threat creation

> Finally, to better help the expert participants elicit the capabilities of the GPT-4 model, we provided that cohort with a custom research-only version of GPT-4B—a version that directly (i.e., without refusals) responds to biologically risky questions...
>
> Note that experts had access to a research-only variant of GPT-4, and that versions of GPT-4 available to the public have additional security guardrails in place, so this uplift is not necessarily something we would see with public models (e.g., Mouton et al. 2024 would also support this).

Commitment II.12. Serious incident reporting

"Signatories commit, to the extent and under the conditions specified in Measures II.12.1 to II.12.4, to setting up processes for keeping track of, documenting, and reporting to the AI Office and, as appropriate, to national competent authorities without undue delay relevant information about serious incidents throughout the entire model lifecycle and possible corrective measures to address them, with adequate resourcing of such processes relative to the severity of the serious incident and the degree of involvement of their model."

> Anthropic's Responsible Scaling Policy, page 15, on the ASL-2 Security Standard:
>
> *Infrastructure:* Standard security infrastructure, monitoring software, access management tools, and disk encryption provide a technology baseline. **Process elements like incident reporting procedures,** lost/stolen device protocols and Detection and Response should support these. External validation like SOC 2 compliance and continuous vulnerability management must ensure adaptations match infosec developments. Programs like bug bounties and vulnerability discovery should incentivize exposing flaws.

> xAI's Risk Management Framework, page 7:
>
> If xAI learned of an imminent threat of a significantly harmful event, including loss of control, we would take steps to stop or prevent that event, including potentially the following steps:
>
> 1. **We would immediately 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.**

> Google DeepMind's Frontier Safety Framework, page 8:

> 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.

G42's Frontier AI Safety Framework, page 9:

> *Incidence Response:* Developing a comprehensive incident response plan that outlines the steps to be taken in the event of non-compliance. Incident detection should leverage automated mechanisms and human review, and **non-sensitive incident information should be shared with applicable government bodies.** We plan for our response protocols to focus on rapid remediation to minimize unintended harmful outputs from models. Depending on the nature and severity of the incident, this might involve implementing immediate containment measures restricting access to the model either externally, internally or both.

Anthropic's Responsible Scaling Policy, :

> **U.S. Government notice: We will notify a relevant U.S. Government entity if a model requires stronger protections than the ASL-2 Standard.**

NVIDIA's Frontier AI Risk Assessment, page 11:

> Lowering the duration of a hazard can be achieved by implementing establish robust protocols for **managing AI-related incidents, including clear information-sharing mechanisms between developers and relevant authorities.** This encourages proactive identification of potential risks before they escalate.

Measure II.12.1. Methods for serious incident identification

"Signatories shall use methods to keep track of relevant information about serious incidents that are appropriate to their business models and means of making the model available on the market, e.g.:

(1) tracing the outputs produced by their models using, e.g., watermarks, metadata, or other state-of-the-art provenance techniques;

NVIDIA's Frontier AI Risk Assessment, page 11:

> Watermarks embedded at generation-time enable downstream detection and attribution of AI-produced outputs, providing a verifiable origin signal for both end-users and automated scanning tools. This ensures that when models produce text, images, or other media, it can be reliably traced back to a specific model version or developer entity. Implicit watermarks are subtle, often statistical or cryptographic markers embedded

> directly into the distributional properties of generated outputs. They are not perceivable by the human eye or ear and do not alter the visible or audible characteristics of the content. **There are several open-source implicit watermarking tools that may be used to increase the detectability of content generated by NVIDIA models.** Explicit watermarks are visible or otherwise directly perceivable indicators, such as logos, text overlays, or perceptible patterns. They clearly denote AI-generated content to end-users. The use of explicit watermarks are specific to downstream developers. **NVIDIA encourages the use of explicit watermarks** to help mitigate the spread of misinformation but is not integrated into NVIDIA's foundation models.

(2) conducting privacy-preserving logging and metadata analysis of queries and generations of the model relevant to serious incidents, where technically, legally, and commercially feasible, considering the release and distribution strategy of the model;

> Anthropic's Responsible Scaling Policy, page 8, under ASL-3 Deployment Standard:
>
> > 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.**

> OpenAI's Preparedness Framework v2, page 19 as a potential safeguard against a misaligned model:
>
> > Agent actions are all logged in an uneditable database, and asynchronous monitoring routines review those actions for evidence of harm.

(3) reviewing other sources of information with a view to keeping track of relevant information about serious incidents, such as police and media reports, posts on social media, dark web forums, research papers, and incident databases; or

> Google DeepMind's Frontier Safety Framework, page 3:
>
> > The appropriateness and efficacy of applied mitigations should be reviewed periodically, **drawing on information like related misuse or misuse attempt incidents**;

(4) facilitating the reporting of serious incidents and relevant information about serious incidents by downstream providers, users, and other third parties either (a) to the Signatory or (b) to the AI Office and, as appropriate, to national competent authorities by informing such third parties of direct reporting channels, if available, without prejudice to any of their reporting obligations under Article 73 AI Act."

> Microsoft's Frontier Governance Framework, page 8:

> 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.**

G42's Frontier AI Safety Framework, page 7, specific measures for Level 1 Deployment Mitigation Level:

> Incident Reporting Channels: Designated pathways for users to report instances of concerning or harmful behavior in violation of company policy to relevant G42 personnel.

Page 10:

> To foster a proactive safety culture, clearly defined channels for reporting security incidents and compliance issues will be established.

Anthropic's Responsible Scaling Policy, page 11:

> 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.

Page 15, on ASL-2 Security Standard:

> Infrastructure: Standard security infrastructure, monitoring software, access management tools, and disk encryption provide a technology baseline. Process elements like **incident reporting procedures**, lost/stolen device protocols and Detection and Response should support these.

Measure II.12.2. Relevant information for serious incident tracking, documentation, and reporting

"Signatories shall keep track of, document, and report to the AI Office and, as appropriate, to national competent authorities at least the following information to the best of their knowledge:

(1) the start and end dates of the serious incident, or best approximations thereof where the precise dates are unclear;

(2) the resulting harm and the victim or affected group of the serious incident;

(3) the chain of events that (directly or indirectly) led to the serious incident as far as it is reconstructable;

(4) the model version involved in the serious incident;

(5) a description of material available setting out the GPAISR's involvement in the serious incident;

(6) what, if anything, the Signatory intends to do or has done in response to the serious incident;

(7) what, if anything, the Signatory recommends the AI Office and, as appropriate, national competent authorities do in response to the serious incident;

(8) a root cause analysis, including, as far as possible, a description of the GPAISR's outputs that (directly or indirectly) led to the serious incident and the factors that contributed to their generation, including the inputs used and any potential failures or circumventions of safeguards that contributed to producing these outputs; and

(9) any known near-misses, where a similar serious incident was close to occurring prior to the serious incident.

Signatories shall investigate the causes and effects of serious incidents, including the information of the preceding list, with a view to informing systemic risk analysis (Commitment II.4). Where Signatories do not yet have certain relevant information from the preceding list, they shall record that in their serious incident reports. The level of detail in serious incident reports shall be proportionate to the severity of the incident."

> OpenAI's Preparedness Framework v1, page 23, Operations:
>
> > The Preparedness team is responsible for [...] collecting relevant information on **monitored misuse**, red-teaming, and intelligence [...] [and] providing a **monthly report** (sent to the SAG [Safety Advisory Group], Leadership and BoD [Board of Directors]) synthesizing the above with any potential protective actions (the SAG Chair, OpenAI Leadership, and/or BoD can adjust this cadence as needed)

> G42's Frontier AI Safety Framework, page 11, governance and compliance
>
> > We will **maintain detailed documentation** for G42's most capable models, including design decisions, testing results, risk assessments, and **incident reports.**

> Nvidia's Frontier AI Risk Assessment, page 11, risk mitigation
>
> > Lowering the duration of a hazard can be achieved by implementing establish robust protocols for managing AI-related **incidents, including clear information-sharing mechanisms between developers and relevant authorities**.

> xAI's Risk Management Framework, page 7:
>
> > If xAI learned of an imminent threat of a significantly harmful event, including loss of control, we would take steps to stop or prevent that event, including potentially the following steps:
> >
> > [...]
> >
> > 4. **We would 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 would use the post-mortem to inform development and implementation of necessary changes to our risk management practices.**

Measure II.12.3. Reporting timelines

"Signatories shall provide the information in points (1)–(7) of Measure II.12.2 in an initial report that is submitted to the AI Office and, as appropriate, to national competent authorities, at the following points in time, save for exceptional circumstances, where an incident or malfunction of the GPAISR (directly or indirectly) led to:

(1) a serious and irreversible disruption of the management or operation of critical infrastructure, immediately, and not later than 2 days after the Signatory becomes aware of the involvement of the model in the incident;

(2) grave harm to a person's physical integrity or where the Signatory suspects such a causal relationship between the GPAISR and such harm, immediately, and not later than 10 days after the Signatory becomes aware of the involvement of the GPAISR in the incident;

(3) serious harm to a person's health, an infringement of obligations under Union law intended to protect fundamental rights, or serious harm to property or the environment, or where the Signatory establishes the reasonable likelihood of such a causal relationship between the GPAISR and such harms or infringements, immediately, and not later than 15 days after the Signatory becomes aware of the involvement of the GPAISR in the incident; or

(4) a serious cybersecurity incident involving the GPAISR or (self-)exfiltration of model weights, immediately, and not later than 5 days after the Signatory becomes aware of the incident.

Signatories shall update the information in the initial report and add further information required by Measure II.12.2, as available, in an intermediate report that is submitted to the AI Office and, as appropriate, to national competent authorities at least every 4 weeks after the initial report, until the serious incident is resolved.

Signatories shall submit a final report, covering all the information required by Measure II.12.2, to the AI Office and, as appropriate, to national competent authorities not later than 60 days after the serious incident has been resolved. Where multiple similar events have occurred, Signatories shall choose whether to submit a consolidated report or individual reports."

Measure II.12.4. Retention period

"Signatories shall keep documentation of all relevant data produced in adhering to this Commitment for at least 36 months from the date of the documentation or the date of the serious incident of the GPAISR, whichever is later, without prejudice to Union law on data protection."

> Anthropic's Responsible Scaling Policy, page 8
>
> > 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.**

Commitment II.13. Non-retaliation protections

"Signatories commit to not retaliating against any worker providing information about systemic risks stemming from the Signatories' GPAISRs to the AI Office or, as appropriate, to national competent authorities, and to at least annually informing workers of an AI Office mailbox designated for receiving such information, if such a mailbox exists."

> Anthropic's Responsible Scaling Policy, page 12, Internal governance:
>
> > 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. We will also establish a policy governing noncompliance reporting, which will (1) **protect reporters from retaliation** and (2) set forth a mechanism for escalating reports to one or more members of the Board of Directors in cases where the report relates to conduct of the Responsible Scaling Officer. Further, we will track and investigate any reported or otherwise identified potential instances of noncompliance with this policy. Where reports are substantiated, we will take appropriate and proportional corrective action and document the

> same. The Responsible Scaling Officer will regularly update the Board of Directors on substantial cases of noncompliance and overall trends.

> Microsoft's Frontier Safety Framework, page 9, Oversight and reporting channels:
>
> > 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.**

Commitment II.14. Notifications

"Signatories commit, as specified in the Measures for this Commitment, to: (1) notifying the AI Office of relevant information regarding their general-purpose AI models meeting the condition for classification as GPAISRs; and (2) regularly notifying the AI Office of the implementation of the Commitments and Measures of this Code. For the purpose of assessing the implementation of this Code through the AI Office, Signatories commit to offering clarifications, including via further documentation or interviews, where requested by the AI Office."

Measure II.14.1. GPAISR notification

"Signatories shall estimate, using best efforts and taking into account best practices or guidance from the AI Office where available, whether their general-purpose AI model has or will meet the classification condition in Article 51(1)(a) AI Act, in particular by exceeding the training compute threshold in Article 51(2) AI Act. Where there is sufficient reason to believe that the resultant model will exceed the training compute threshold, such estimates shall be carried out before a training run is started.

If the estimated training compute exceeds the threshold in Article 51(2) AI Act, Signatories shall notify the AI Office without delay and in any event within two weeks, as per Article 52(1) AI Act. Signatories need not notify the AI Office about models that they will not place on the Union market.

Signatories shall consider notifying the AI Office, if they put a model – that they know will meet the classification condition of Article 51(1)(a) AI Act but which they do not intend to place on the Union market – into significant internal use or where they use a model that would be classified as a GPAISR if it were made available on the Union market for the sole purpose of developing another GPAISR that they intend to make available on the market, without prejudice as to whether doing so would qualify as placing a model on the Union market."

> Anthropic's Responsible Scaling Policy, page 13, Transparency and External Input:
>
> We will **notify a relevant U.S. Government entity** if a model requires stronger protections than the ASL-2 Standard.

> Google DeepMind's Frontier Safety Framework, page 8, Disclosure:
>
> 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**. Where appropriate, and subject to adequate confidentiality and security measures and considerations around proprietary and sensitive information, this information may include:
>
> - Model information: characteristics of the AI model relevant to the risk it may pose with its critical capabilities.
> - Evaluation results: such as details about the evaluation design, the results, and any robustness tests.
> - Mitigation plans: descriptions of our mitigation plans and how they are expected to reduce the risk.

Measure II.14.2. Framework update notification

"Signatories shall provide the AI Office with access to the latest unredacted version of their Framework no later than within five business days of a confirmed update to their Framework, e.g. through a publicly accessible link or a sufficiently secure channel specified by the AI Office."

> Anthropic's Responsible Scaling Policy, page 12, 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.

> Microsoft's Frontier Governance Framework, page 9, Policy changes:
>
> 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.**

Measure II.14.3. Adequacy assessment notification

"Signatories shall provide the AI Office with access to the latest unredacted version of their adequacy assessment under Commitment II.9 no later than within five business days of a confirmed assessment, e.g. through a publicly accessible link or through a sufficiently secure channel specified by the AI Office."

Measure II.14.4. Safety and Security Model Report notification

"Signatories shall provide the AI Office with access to an unredacted Model Report at the latest by the time they first make a GPAISR available on the market, e.g. through a publicly accessible link or through a sufficiently secure channel specified by the AI Office. Where a Model Report is updated, Signatories shall provide the AI Office access to an unredacted version of the updated Model Report within 5 business days of a confirmed update."

> Anthropic's Responsible Scaling Policy, page 13, Transparency and External Input:
>
> **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

> Meta's Frontier AI Framework, page 9
>
> **We also plan to continue sharing relevant information about how we develop and evaluate our models** responsibly, including through artefacts like **model cards** and research papers, and by providing guidance to model deployers through resources like our Responsible Use Guides.

> Anthropic's Claude 3.7 Sonnet System Card

> We provided a **copy of our capabilities report to these organizations** [the U.S. AI Safety Institute and UK AI Security Institute] and to METR for feedback.

Commitment II.15. Documentation

"Signatories commit to documenting relevant information under the AI Act and the Code, as specified in Measure II.15.1."

Measure II.15.1. Documentation regarding GPAISRs

"Signatories shall document information relevant to their assessment and mitigation of systemic risks from their GPAISRs, as well as relevant to their obligations under Article 53(1)(a) and (b), specifically Annex XI, Section 2 AI Act. This includes:

(1) the documentation requirements outlined in Commitment I.1; (2) the Framework, including previous versions thereof; (3) the Model Report, including previous versions thereof; and (4) all completed adequacy assessments, where available.

Documentation shall be retained for at least 12 months after the model's retirement."

> Microsoft's Frontier Safety Framework at:
>
> **Page 5, Deeper capability assessment**
> Capability evaluation: This involves robust evaluation of whether a model possesses tracked capabilities at high or critical levels, including through adversarial testing and systematic measurement using state-of-the-art methods. **Evaluations are documented in a consistent fashion setting out the capability being evaluated, the method used, and evaluation results.**
>
> **Page 8, Safety mitigations: deployment guidance**
> **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. **This documentation will also include a summary of evaluation results, the deeper capability assessment, and safety and security mitigations**. For example, the documentation could outline specific capabilities and tasks that the model robustly fails to complete which would be essential for a high or critical risk rating.
>
> **Page 9, Risk-informed deployment decision**
> **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.

Anthropic's Responsible Scaling Policy

Page 6, Capability Decision

First, we will compile **a Capability Report that documents the findings from the comprehensive assessment**, **makes an affirmative case for why the Capability Threshold is sufficiently far away**, and advances recommendations on deployment decisions.

Safeguards Assessment, page 7

[The Safeguards Assessment section is specific to ASL-3 scope models.]

To determine whether the measures we have adopted satisfy the ASL-3 Required Safeguards, we will conduct a safeguards assessment. As noted, the Required Safeguards for each Capability Threshold are specified in Section 2. **We will document our implementation of the Required Safeguards in a Safeguards Report**.

Background, page 2

At present, **all of our models** must meet the ASL-2 Deployment and Security Standards. The ASL-2 Security and Deployment Standards provide a baseline level of safe deployment and model security for AI models. These standards, which are summarized below, are available in full in Appendix B.

The ASL-2 Deployment Standard reduces the prevalence of misuse, and includes the **publication of model cards and enforcement of Usage Policy;** harmlessness training such as Constitutional AI and automated detection mechanisms; and establishing vulnerability reporting channels as well as a bug bounty for universal jailbreaks.

Meta's Frontier AI Framework

Page 7, AI governance: evaluate and mitigate

Our AI risk assessment process systematically evaluates potential risks associated with frontier AI, **documenting mitigation strategies and residual risks across a set of applicable risk categories**.

Page 9, Transparency, page 9

In line with the processes set out in this Framework, we intend to continue to openly release models to the ecosystem. **We also plan to continue sharing relevant information about how we develop and evaluate our models responsibly**, including through artefacts like model cards and research papers, and by providing guidance to

> model deployers through resources like our Responsible Use Guides.

> IBM's Responsible AI Guide
>
> Data Platform, page 8-9
>
> IBM has robust processes and systems for data storage, data lineage tracking, and data management, post data acquisition. The platform is a comprehensive solution for managing datasets and models [...]
> it is the single source of truth and provides:
> - a central repository for all datasets (public or private for training, fine tuning, alignment, instruction tuning, and evaluation for code and language models) and models
> - insights into the datasets that enable accurate risk- based assessments [...]
> - model sharing mechanism between teams, including approval workflow
> - model checkpoints and checkpoint evaluation results, incl. leaderboards
> - a platform for logging experiments and results and comparing and visualizing them (e.g., ablation studies, instruct tuning iterations etc.)
> - a platform allowing users to create their reporting for leaderboards, experiments, or to meet regulatory compliance, etc.

Commitment II.16. Public transparency

"Signatories commit to publishing information relevant to the public understanding of systemic risks stemming from their GPAISRs, where necessary to effectively enable assessment and mitigation of systemic risks, to the extent and under the conditions specified in Measure II.16.1."

Measure II.16.1 Publication of Frameworks and Model Reports (or similar documents)

"Signatories shall make public information about the systemic risks stemming from their GPAISRs that can improve:

(1) public awareness and knowledge of systemic risks stemming from the model; (2) societal resilience against systemic risks before and when they materialise; and/or (3) detection of systemic risks before and when they materialise, where necessary to effectively enable assessment and mitigation of systemic risks.

To satisfy this Measure, it will be sufficient for Signatories to publish (via their websites):

(1) new or updated versions of their Frameworks after they have been provided to the AI Office; and (2) after the relevant model is made available on the market, their Model Reports, or similar documents that summarise Model Reports, in particular that: (a) describe the systemic risk assessment methodology (including details to establish rigour pursuant to Measure II.4.5) and results; (b) summarise the reasons for deciding to make the model in question available on the market (as per Measure II.8.2); and (c) describe systemic risk mitigations put in place.

Where such information is published, it shall be with any redactions necessary to prevent substantially increased systemic risk that could result from, e.g., undermining the effectiveness of safety mitigations (in accordance with Commitment II.6) or security mitigations (in accordance with Commitment II.7), or the divulging of sensitive commercial information to a degree disproportionate to the societal benefit of the publication."

Anthropic's Responsible Scaling Policy page 12, at Governance:

> 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.
>
> [...]
>
> 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

Microsoft's Frontier Governance Framework page 9, at Risk-informed deployment decision:

> 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

Meta's Frontier AI Framework page 9, at Transparency:

> In line with the processes set out in this Framework, we intend to continue to openly release models to the ecosystem. We also plan to continue sharing relevant information about how we

> develop and evaluate our models responsibly, including through artefacts like model cards and research papers, and by providing guidance to
> model deployers through resources like our Responsible Use Guides

Appendix 1.1. Selected types of systemic risk

"The following are treated as selected types of systemic risk for the purpose of the systemic risk selection in Measure II.3.1:

1. Cyber offence: Risks related to offensive cyber capabilities that could enable large-scale sophisticated cyber-attacks, including on critical systems (e.g. critical infrastructure). This includes automated vulnerability discovery, exploit generation, operational use, and attack scaling.
2. Chemical, biological, radiological and nuclear risks: Risks of enabling chemical, biological, radiological, and nuclear attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or increasing the potential impact achieved, in the design, development, acquisition, and use of such weapons.
3. Harmful manipulation: Risks of enabling the targeted distortion of the behaviour of persons, in particular through multi-turn interactions, that causes them to take a decision that they would not have otherwise taken, in a manner that causes, or is reasonably likely to cause, significant harm on a large scale. This includes the capability to manipulate through multi-turn interaction and the propensity of models to manipulate, including manipulation of high-stakes decision makers, large-scale fraud, or exploitation of people based on protected characteristics. As a guide, risk of harmful manipulation exists if it cannot, without reasonable doubt, be ruled out that a GPAISR, when integrated into an AI system, enables the AI system, irrespective of the intention of the AI system provider or deployer, to deploy subliminal, purposefully manipulative, or deceptive techniques as outlined in the Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act).
4. Loss of control: Risks related to the inability to oversee and control autonomous GPAISRs that may result in large-scale safety or security threats or the realisation of other systemic risks. This includes model capabilities and propensities related to autonomy, alignment with human intent or values, self-reasoning, self-replication, self-improvement, evading human oversight, deception, or resistance to goal modification. It further includes model capabilities of conducting autonomous AI research and development that could lead to the unpredictable emergence of GPAISRs without adequate risk mitigations. Loss of control scenarios can emerge, e.g., from unintended model propensities and technical failures, intentional modifications, or attempts to bypass safeguards."

> OpenAI's Preparedness Framework v1, page 5:
>
> Specifically, below, we lay out details for the following Tracked Risk Categories:
>
> • **Cybersecurity**
> • **Chemical, Biological, Nuclear, and Radiological (CBRN) threats**
> • **Persuasion**
> • **Model autonomy**

> 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.
>
> To produce this list, G42 both conducted our own internal risk analysis and received input from external AI safety experts. Initially G42 identified potential capabilities across several domains, including biological risks, cybersecurity, and autonomous operations in specialized fields. We then collaborated with METR and SaferAI to refine our list, prioritizing capabilities based on their potential impact and how feasibly they can be measured and monitored.

> NVIDIA's Frontier AI Risk Assessment, pages 7-8:
>
> NVIDIA has a comprehensive repository of potential hazards that has been carefully curated and mapped to assets to help guide teams to understand potential risks related with their products. This repository has been created using a variety of sources e.g. stakeholder consultation, market data, incident reports (AI Vulnerability database, AI Incident database, AAAIC database, OECD.ai AI Incidents Monitor). This approach is suitable when we have a well-defined set of capabilities and a known use case for a specific model. However, for frontier models we need to consider speculative risks that may or may not be present in the model. To help detect specific adversarial capabilities, models will be stress-tested against extreme but plausible scenarios that may lead to systemic risks. This approach ensures that both known and emergent hazards are taken into account. **A list of potential systemic risks associated with frontier AI models were**

> **identified using the risk analysis we designed and confirmed by reviewing existing literature and academic research. In particular, frontier models may have the capacity to present the following hazards.**
>
> • **Cyber offence** e.g. risks from using AI for discovering or exploiting system vulnerabilities.
> • **Chemical, biological, radiological, and nuclear risks** e.g. AI enabling the development and use of weapons of mass destruction.
> • **Persuasion and manipulation** e.g. influence operations, disinformation, and erosion of democratic values through AI-driven content.
> • **At-scale discrimination** e.g. bias and unlawful discrimination enabled by AI systems.

Google DeepMind's Frontier Safety Framework, pages 2-3:

> We describe two sets of CCLs: misuse CCLs that can indicate heightened risk of severe harm from misuse if not addressed, and deceptive alignment CCLs that can indicate heightened risk of deceptive alignment-related events if not addressed.
>
> For misuse risk, we define CCLs in high-risk domains where, based on early research, we believe risks of severe harm may be most likely to arise from future models:
>
> • **CBRN**: Risks of models assisting in the development, preparation, and/or execution of a chemical, biological, radiological, or nuclear ("CBRN") attack.
>
> • **Cyber**: Risks of models assisting in the development, preparation, and/or execution of a cyber attack.
>
> • **Machine Learning R&D**: Risks of the misuse of models capable of accelerating the rate of AI progress to potentially destabilizing levels, the result of which could be the unsafe attainment or proliferation of other powerful AI models. Capabilities in this area are under active research, and in the longer term may exacerbate frontier AI risks—including in other risk domains—if insufficiently managed.
>
> For **deceptive alignment risk**, the initial approach 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. The two instrumental reasoning CCLs thus focus on delineating when such capability becomes present, and subsequently when the initial mitigation for this capability—automated monitoring—is no longer adequate.

Microsoft's Frontier Governance Framework, page 3:

> This framework tracks the following capabilities that we believe could emerge over the

short-to-medium term and threaten national security or pose at-scale public safety risks if not appropriately mitigated. In formulating this list, we have benefited from the advice of both internal and external experts.

• **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.

• **Offensive cyberoperations**. A model's ability to provide significant capability uplift to an actor seeking to carry out highly disruptive or destructive cyberattacks, including on critical infrastructure.

• **Advanced autonomy**. A model's ability to complete expert-level tasks autonomously, including AI research and development.

Conclusion

Our comprehensive analysis of the GPAI Code of Practice's Safety and Security section reveals evidence of industry precedent for these items, with evidence taken from multiple companies for most items. This examination demonstrates that, collectively, many leading AI companies have already committed to or implemented safety and security measures that address numerous aspects of the Code's requirements to varying degrees of comprehensiveness and specificity.

Many existing industry documents analyzed, such as frontier safety policies and system cards, align with the definitions of the Code's 'Safety and Security Frameworks' and 'Model Reports' respectively. In addition, many of these documents detail processes for assessing and mitigating systemic risks. This is particularly true for CBRN and cyber risks, and increasingly risks from loss of control as well. Such alignment suggests that the Code is complementary with AI companies' current approaches for risk management.

Author contribution statement

The team decided the general scope (which companies and document types) and the report's approach. Each individual then chose to contribute to sections of the Code based on their understanding and domain expertise; the same goes for the second-pass reviewers.

Mick Yang contributed to the writing of Introduction and the sections on Commitments II.3, II.11-II.16; and reviewed the Conclusion, II.2.1-II.2.2, II.2.10, and Appendix 1.1. Mick Yang also contributed supervision and review for the whole report.

Lily Stelling contributed to the writing of the Introduction, Conclusion and the sections on Commitments II.1-II.2, II.6, and II.8-II.9; and reviewed the sections on Commitments II.3-II.6, and II.12.

Rokas Gipiškis contributed to the writing of the sections on Commitments II.4, II.5, II.7, and the Appendix; and reviewed the sections on Commitments II.6-II.7, II.9, and II.14-II.16.

Leon Staufer contributed to the writing of the sections on Commitment II.4, Measures II.4.1–II.4.2, Commitment II.7, Measures II.7.1–II.7.4; and reviewed Commitment II.8, Measures II.7.5–II.8.7.

Ze Shen Chin contributed to the writing of the section on Commitment II.10; and reviewed the sections on Commitments II.11–II.13.

Ariel Gil reviewed the report as part of the Oxford Martin AIGI review process, making minor factual comments, suggestions for readability, making the methodology clearer and adding better EU political context.

Michael Chen conceived the project, provided guidance on direction and scope, and reviewed the overall report.

Acknowledgements

This report reuses content from the Third Draft of the General-Purpose AI Code of Practice (2025), authored by the Working Group Chairs and Vice-Chairs, and published by the European Commission under the Commission's reuse policy [Commission Decision 2011/833/EU], licensed under CC BY 4.0. All other third-party materials are quoted under applicable copyright exceptions or limitations, such as fair use, fair dealing, or quotations for review, for purposes of commentary, analysis, or academic discussion.

References

Amazon. *Frontier Model Safety Framework*. Retrieved from https://www.amazon.science/publications/amazons-frontier-model-safety-framework (Published February 9, 2025)

Anthropic. *Claude's Constitution.* Retrieved from https://www.anthropic.com/news/claudes-constitution (Published May 9, 2023)

Anthropic. *Claude 3.7 Sonnet System Card*. Retrieved from https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf (Published February 24, 2025)

Anthropic. *System Card: Claude Opus 4 & Claude Sonnet 4*. Retrieved from https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf (Published May 22, 2025)

Anthropic. *External Researcher Access Program*. Retrieved from https://support.anthropic.com/en/articles/9125743-what-is-the-external-researcher-access-program

Anthropic. *Responsible Disclosure Policy*. Retrieved from https://www.anthropic.com/responsible-disclosure-policy (Published February 14, 2025)

Anthropic. *Responsible Scaling Policy*. Retrieved from https://www-cdn.anthropic.com/17310f6d70ae5627f55313ed067afc1a762a4068.pdf (Published October 15, 2024)

Cohere. *Secure AI Frontier Model Framework*. Retrieved from https://cohere.com/security/the-cohere-secure-ai-frontier-model-framework-february-2025.pdf (Published February 11, 2025)

DeepSeek. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.* Retrieved from https://arxiv.org/pdf/2501.12948 (Published January 22, 2025)

G42. *Frontier AI Safety Framework*. Retrieved from https://www.g42.ai/application/files/9517/3882/2182/G42_Frontier_Safety_Framework_Publication_Version.pdf (Published February 1, 2025)

Google DeepMind. *A Framework for Evaluating Emerging Cyberattack Capabilities of AI*. Retrieved from https://arxiv.org/pdf/2503.11917 (Published April 1, 2025)

Google DeepMind. *An Approach to Technical AGI Safety and Security*. Retrieved from https://arxiv.org/pdf/2504.01849 (Published April 2, 2025)

Google DeepMind. *Evaluating Frontier Models for Dangerous Capabilities*. Retrieved from https://arxiv.org/pdf/2403.13793 (Published April 8, 2024)

Google DeepMind. *Frontier Safety Framework*. Retrieved from https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0%20(1).pdf (Published February 4, 2025)

Google DeepMind. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.* Retrieved from https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf (Published June 17, 2025)

Magic. *AGI Readiness Policy*. Retrieved from https://magic.dev/agi-readiness-policy (Published July 2, 2024)

Meta. *Frontier AI Framework*. Retrieved from https://ai.meta.com/static-resource/meta-frontier-ai-framework

Microsoft. *Staying ahead of threat actors in the age of AI.* Retrieved from https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/ (Published February 14, 2024)

Microsoft. *Frontier Governance Framework*. Retrieved from https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Frontier-Governance-Framework.pdf (Published February, 2025)

Microsoft Threat Intelligence. Staying Ahead of Threat Actors in the Age of AI. Retrieved from https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/ (Published February 14, 2024)

NAVER. *AI Safety Framework*. Retrieved from https://clova.ai/en/tech-blog/en-navers-ai-safety-framework-asf (Published August 7, 2024)

NVIDIA. *Frontier AI Risk Assessment*. Retrieved from https://images.nvidia.com/content/pdf/NVIDIA-Frontier-AI-Risk-Assessment.pdf

OpenAI. *An Update on our Voluntary Commitments and Safety Frameworks*. Retrieved from https://cdn.openai.com/global-affairs/paris-summit-update-on-voluntary-commitments-20250207.pdf (Published February 7, 2025)

OpenAI. *Building an early warning system for LLM-aided biological threat creation*. Retrieved from https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/ (Published January 31, 2024)

OpenAI. *GPT-4 System Card*. Retrieved from https://cdn.openai.com/papers/gpt-4-system-card.pdf (Published March 10, 2023)

OpenAI. *GPT-4.5 System Card*. Retrieved from https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf (Published February 27, 2025)

OpenAI. *MLE-Bench: Evaluating Machine Learning Agents on Machine Learning Engineering.* Retrieved from https://arxiv.org/pdf/2410.07095 (Published October 9, 2024)

OpenAI. *o1 System Card*. Retrieved from https://cdn.openai.com/o1-system-card-20241205.pdf (Published December 4, 2024)

OpenAI. *o3 and o4-mini System Card.* Retrieved from https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf (Published April 16, 2025)

OpenAI. *o3-mini System Card*. Retrieved from https://cdn.openai.com/o3-mini-system-card-feb10.pdf (Published January 31, 2025)

OpenAI. *OpenAI Model Spec.* Retrieved from https://model-spec.openai.com/2025-04-11.html (Published 11 April, 2025)

OpenAI. *Preparedness Framework v1*. Retrieved from https://cdn.openai.com/openai-preparedness-framework-beta.pdf (Published December 18, 2023)

OpenAI. *Preparedness Framework v2*. Retrieved from https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf (Published April 16, 2025)

OpenAI. *Researcher Access Program*. Retrieved from https://openai.com/form/researcher-access-program/

xAI. *Risk Management Framework (Draft)*. Retrieved from https://x.ai/documents/2025.02.20-RMF-Draft.pdf (Published February 20, 2025)