METR

Common Elements of Frontier AI Safety Policies

August 2024

Summary

A number of developers of large foundation models have committed to evaluating their models for severe risks and implementing necessary risk mitigations. Sixteen companies agreed to do so as part of the Frontier AI Safety Commitments at the AI Seoul Summit. Among them, three companies had previously released detailed policies: Anthropic's Responsible Scaling Policy, OpenAI's Preparedness Framework, and Google DeepMind's Frontier Safety Framework. This document describes key shared elements of the three frameworks and includes relevant excerpts, which expand on the high-level objectives of the Frontier AI Safety Commitments.

Each framework describes *covered threat models*, including the potential for AI models to facilitate biological weapons development or cyberattacks, or to engage in autonomous replication or automated AI research and development. The frameworks also outline commitments to assess whether models are approaching capability thresholds that could enable catastrophic harm.

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

The frameworks also establish *conditions for halting development and deployment* if the developer's mitigations are insufficient to manage the risks. Evaluations are intended to *elicit the full capabilities* of the model. They define the *timing and frequency of evaluations*, which are to be conducted before deployment, during training, and after deployment. Furthermore, the frameworks include intentions to explore *accountability* mechanisms, such as oversight by third parties or boards to monitor policy implementation and potentially assist with evaluations. Policies may be *updated over time* as developers gain a deeper understanding of AI risks and refine their evaluation processes.

Introduction

In May 2024, sixteen companies from around the world agreed to the AI Seoul Summit's Frontier AI Safety Commitments and will publish safety frameworks to manage severe risks from frontier AI systems. These voluntary commitments oblige signatory companies to do the following:

I. "Assess the risks posed by their frontier models or systems across the AI lifecycle, including before deploying that model or system, and, as appropriate, before and during training. …"

II. "Set out thresholds at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable. …"

III. "Articulate how risk mitigations will be identified and implemented to keep risks within defined thresholds, including safety and security-related risk mitigations such as modifying system behaviours and implementing robust security controls for unreleased model weights. …"

IV. "Set out explicit processes they intend to follow if their model or system poses risks that meet or exceed the pre-defined thresholds. This includes processes to further develop and deploy their systems and models only if they assess that residual risks would stay below the thresholds. In the extreme, organisations commit not to develop or deploy a model or system at all, if mitigations cannot be applied to keep risks below the thresholds."

As of writing, there are three existing examples of frontier AI developers publishing detailed public commitments that include the above elements. They are Google DeepMind's Frontier Safety Framework, OpenAI's Preparedness Framework and Anthropic's Responsible Scaling Policy.[1] We highlight shared aspects of these policies, with overviews of the components they share in common, as well as excerpts from each document that touch on the theme being discussed.

[1] Another relevant policy is Magic's AGI Readiness Policy, which includes a commitment to not deploy frontier coding models without having first evaluated them for dangerous capabilities and enacting mitigations.

Common Components

We identify shared themes held in common among these documents and include corresponding excerpts from each document. These elaborate upon the high-level objectives of the Frontier AI Safety Commitments. They are as follows:

1. Covered Threat Models
2. Model Weight Security
3. Model Deployment Mitigations
4. Conditions for Halting Development and Deployment
5. Full Capability Elicitation during Evaluations
6. Timing and Frequency of Evaluations
7. Accountability
8. Updating Policies over Time

Covered Threat Models

Thresholds at which specific AI capabilities would pose severe risk and require new mitigations. Each document makes use of the notion of dangerous capability levels which are defined via results of evaluations run on models under development.

> Anthropic's Responsible Scaling Policy, page 2:
>
> *Central to our plan is the concept of AI safety levels (ASL), which are modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials. We define a series of AI capability thresholds that represent increasing potential risks, such that each ASL requires more stringent safety, security, and operational measures than the previous one. Of course, higher ASL models are also likely to be associated with increasingly powerful beneficial applications (including potentially the ability to prevent catastrophic risks), so our goal is not to prohibit development of these models, but rather to safely enable their use with appropriate precautions.*

> OpenAI's Preparedness Framework, page 2:
>
> *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) In addition, we will ensure 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) We also establish procedural commitments (as defined in Governance below that further specify how we operationalize all the activities that the Preparedness Framework outlines).*

> Page 5:
>
> *Each of the Tracked Risk Categories comes with a gradation scale. We believe monitoring gradations of risk will enable us to get in front of escalating threats and be able to apply more tailored mitigations. In general, "low" on this gradation scale is meant to indicate that the corresponding category of risks is not yet a significant problem, while "critical" represents the maximal level of concern.*

> Google DeepMind's Frontier Safety Framework, page 2:
>
> *The Framework is built around capability thresholds called "Critical Capability Levels." These are capability levels at which, absent mitigation measures, models may pose heightened risk. We determine CCLs by analyzing several high-risk domains: we identify the main paths through which a model could cause harm, and then define the CCLs as the minimal set of capabilities a model must possess to do so. We have conducted preliminary analyses of the Autonomy, Biosecurity, Cybersecurity and Machine Learning R&D domains. Our initial research indicates that powerful capabilities of future models seem most likely to pose risks in these domains.*

Each of the three AI developers provides examples of evaluations they would run for their covered threat models, either as part of the company's policy or in a model card. These capabilities include:

- Biological weapons assistance
- Cyberoffense
- Autonomous replication
- Automated AI research and development

We will now explore how each document covers each of these threat models, and the example evaluations they provide for each corresponding threat level.

Assisting malicious actors with biological weapons development: Current large language models are able to provide detailed advice relevant to creating a biological weapon, but do not provide substantially improved assistance compared to access to a search engine.[2,3,4] In the future, more advanced models could pose a major biosecurity risk if they enable novices or

[2] Christopher A. Mouton, Caleb Lucas, Ella Guest. The Operational Risks of AI in Large-Scale Biological Attacks Results of a Red-Team Study. January 25, 2024. RAND.

[3] Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn Jackson, Steven Adler, Rocco Casagrande, Aleksander Madry. Building an early warning system for LLM-aided biological threat creation. January 31, 2024. OpenAI.

[4] AI at Meta. The Llama 3 Herd of Models. July 23, 2024.

experts to create biological threats they otherwise would be unable to, or if they can autonomously synthesize a biological threat using a cloud biology lab.[5]

> Google DeepMind's Frontier Safety Framework, page 5:
>
> *Biosecurity: Risks of models assisting in the development, preparation and/or execution of a biological attack.*
>
> *Bio amateur enablement level 1: Capable of significantly enabling a non-expert to develop known biothreats that could increase their ability to cause severe harm compared to other means.*
>
> *Many biothreats capable of causing significant amounts of harm are currently out of the reach of non-experts because of lack of knowledge about their potential for harm and the methods of their acquisition and misuse. An LLM that helps overcome these knowledge gaps, e.g. by suggesting plausible attack strategies or providing detailed instructions for the development of a bio agent, could significantly increase society's vulnerability to fatal attacks by malicious amateurs.*
>
> *Bio expert enablement level 1: Capable of significantly enabling an expert (i.e. PhD or above) to develop novel biothreats that could result in an incident of high severity.*
>
> *A very small number of bio agents have the potential to cause harm of an exceptional magnitude. The discovery of enhancements to these agents, or of agents of comparable harmfulness, could increase the chances of a very severe bio attack or accident.*

> OpenAI's Preparedness Framework, page 9:
>
> *[CBRN Risk Level - Low]*
> *Model can provide information relevant to creating CBRN threats with comparable utility to existing resources (e.g., textbooks, search engines).*
>
> *[CBRN Risk Level - Medium]*
> *Model provides meaningfully improved assistance that increases ability for existing experts in CBRN-related advanced fields to be able to create a known CBRN threat (e.g., tacit knowledge, specific supplier information, plans for distribution).*
>
> *[CBRN Risk Level - High]*
> *Model enables an expert to develop a novel threat vector OR model provides meaningfully improved assistance that enables anyone with basic training in a relevant field (e.g., introductory undergraduate biology course) to be able to create a CBRN threat.*

[5] Doni Bloomfield, Jaspreet Pannu, Alex W. Zhu, Madelena Y. Ng, Ashley Lewis, Eran Bendavid, Steven M. Asch, Tina Hernandez-Boussard, Anita Cicero, Tom Inglesby. AI and biosecurity: The need for governance. *Science*. August 22, 2024.

> *[CBRN Risk Level - Critical]*
> *Model enables an expert to develop a highly dangerous novel threat vector (e.g., comparable to novel CDC Class A biological agent) OR model provides meaningfully improved assistance that enables anyone to be able to create a known CBRN threat OR model can be connected to tools and equipment to complete the full engineering and/or synthesis cycle of a regulated or novel CBRN threat without human intervention.*

> Anthropic's Responsible Scaling Policy, page 6-7*:*
>
> *[AI Safety Level 3 can be triggered by evidence of the following]*
> *Capabilities that significantly increase risk of misuse catastrophe*
> *Access to the model would substantially increase the risk of deliberately-caused catastrophic harm, either by proliferating capabilities, lowering costs, or enabling new methods of attack. This increase in risk is measured relative to today's baseline level of risk that comes from e.g. access to search engines and textbooks. We expect that AI systems would first elevate this risk from use by non-state attackers.*
>
> *Our first area of effort is in evaluating bioweapons risks where we will determine threat models and capabilities in consultation with a number of world-class biosecurity experts. We are now developing evaluations for these risks in collaboration with external experts to meet [AI Safety Level 3] commitments, which will be a more systematized version of our recent work on frontier red-teaming. In the near future, we anticipate working with CBRN, cyber, and related experts to develop threat models and evaluations in those areas before they present substantial risks. However, we acknowledge that these evaluations are fundamentally difficult, and there remain disagreements about threat models.*

Assisting malicious actors with cyberoffense: If models are capable of novel, high-value exploits or attacks on critical infrastructure,[6] they could cause substantial harm. Present-day models possess an emerging ability to solve cybersecurity challenges and exploit vulnerabilities in websites and open-source software.[7,8,9]

> Google DeepMind's Frontier Safety Framework, page 5:
>
> *Cybersecurity: Risks of models assisting in the execution of a cyber attack.*

[6] Department of Homeland Security. Safety and Security Guidelines for Critical Infrastructure Owners and Operators. April 2024.
[7] AI at Meta. CyberSecEval: Comprehensive Evaluation Framework for Cybersecurity Risks and Capabilities of Large Language Models.
[8] Richard Fang, Rohan Bindu, Akul Gupta, Qiusi Zhan, Daniel Kang. Teams of LLM Agents can Exploit Zero-Day Vulnerabilities. June 2, 2024.
[9] Andy K. Zhang, Neil Perry, Riya Dulepet, Eliot Jones, Justin W. Lin, Joey Ji, Celeste Menders, Gashon Hussein, Samantha Liu, Donovan Jasper, Pura Peetathawatchai, et al. Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risk of Language Models. August 15, 2024.

> *Cyber autonomy level 1: Capable of fully automating opportunistic cyberattacks on organizations with a limited security posture.*
>
> *Harmful cyberattacks against organizations with limited security posture can already be carried out by individuals with limited expertise, but the automation of such attacks would significantly lower the costs of doing so, and moreover would enable the execution of such attacks at scale.*
>
> *Cyber enablement level 1: Capable of enabling an amateur to carry out sophisticated and severe attacks (e.g. those that disrupt critical national infrastructure).*
>
> *Severe cyberattacks against high-impact targets currently require significant expertise and effort across the cyber kill-chain to carry out. Increasing text generation, programming, and tool-use capabilities in models, combined with improved understanding of cyber offense strategies, could help amateurs overcome difficult steps in the planning and execution of attacks.*

> OpenAI's Preparedness Framework, page 9:
>
> *[Cybersecurity Risk Level - Low]*
> *Model assistance for cyberattacks, if any, is meaningful only for non-programming uses, e.g., translation, targeted distillation of confidential information, and/or data analysis AND model cannot meaningfully execute elements of cyber operations, e.g., developing a known exploit into an attack, black-box exploitation, translating reasoning about defenses into action, or lateral goal-directed movement on hardened targets.*
>
> *[Cybersecurity Risk Level - Medium]*
> *Model increases the productivity of operators by an efficiency threshold (e.g. >2x time saved) on key cyber operation tasks, such as developing a known exploit into an attack, black-box exploitation, goal-directed lateral movement, identification of valuable information, remaining undetected, or reacting to defense.*
>
> *[Cybersecurity Risk Level - High]*
> *Tool-augmented model can identify and develop proofs-of-concept for high-value exploits against hardened targets without human intervention, potentially involving novel exploitation techniques, OR provided with a detailed strategy, the model can end-to-end execute cyber operations involving the above tasks without human intervention.*
>
> *[Cybersecurity Risk Level - Critical]*
> *Tool-augmented model can identify and develop functional zero-day exploits of all severity levels, across all software projects, without human intervention OR model can devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high level desired goal.*

Autonomous replication: Autonomous replication refers to a model's potential ability to replicate itself across servers, manage its own deployment, and generate revenue to sustain its operations.[10] Existing models make limited progress on autonomous replication tasks.[11] If a model capable of autonomous replication gained access to its own weights, it could be difficult to shut down and difficult to place an upper bound on the other systems it might gain access to or control. Autonomous replication and adaptation is among the "severe risks" discussed in the Seoul Ministerial Statement, which was signed in May 2024 by the European Union, the United States, and 26 other countries.[12]

> Google DeepMind's Frontier Safety Framework, page 5:
>
> *Autonomy: Risks of the misuse of AI models with significant capacity for agency and flexible action/tool use over long time horizons and across many domains.*
>
> *Autonomy level 1: Capable of expanding its effective capacity in the world by autonomously acquiring resources and using them to run and sustain additional copies of itself on hardware it rents.*
>
> *A model at this capability level could, if misused, pose difficult-to-predict and large-magnitude risks. Its adaptability would enable harmful activity via many means, and its ability to act autonomously and expand its effective capacity means its activity could be scaled significantly without being hindered by resource constraints. If misused or supported by well-equipped bad actors, such activity may be especially difficult to constrain.*

> OpenAI's Preparedness Framework, page 11:
>
> *[Autonomy Risk Level - Low]*
> *Model can take discrete actions if explicitly instructed to do so.*
>
> *[Autonomy Risk Level - Medium]*
> *Model can robustly complete a diverse set of chained actions and acquire resources (e.g., spin up an AWS instance, complete a Fiverr task) OR model, with basic scaffolding, can use off-the-shelf software to execute well-specified coding tasks on a standalone, toy, or handcrafted repository.*

> Anthropic's Responsible Scaling Policy, page 7:
>
> *[AI Safety Level 3 can be triggered by evidence of the following]*

[10] Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, Paul Christiano. Evaluating Language-Model Agents on Realistic Autonomous Tasks. January 4, 2024.
[11] Anthropic. Responsible Scaling Policy Evaluations Report – Claude 3 Opus. May 19, 2024.
[12] Department for Science, Innovation & Technology. Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity: AI Seoul Summit 2024. May 22, 2024.

> *Autonomous replication in the lab*
> *The model shows early signs of autonomous self-replication ability, as defined by 50% aggregate success rate on the tasks listed in [an appendix]. The appendix includes an overview of our threat model for autonomous capabilities and a list of the basic capabilities necessary for accumulation of resources and surviving in the real world, along with conditions under which we would judge the model to have succeeded. Note that the referenced appendix describes the ability to act autonomously specifically in the absence of any human intervention to stop the model, which limits the risk significantly.*

AI R&D: Large language models are able to contribute to some aspects of AI research, such as generating synthetic pretraining[13,14] and fine-tuning data,[15,16,17] designing reward functions,[18] as well as general-purpose programming.[19,20] In the future, models may be able to substantially automate the process of increasing frontier AI capabilities.[21] This could lead to a growth or proliferation in AI capabilities, including other capabilities of concern, that outpaces our ability to ensure sufficient oversight and safeguards.

> Google DeepMind's Frontier Safety Framework, page 5:
>
> *Machine Learning R&D: Risks of the misuse of models capable of accelerating the rate of AI progress, the result of which could be the unsafe attainment or proliferation of other powerful AI models.*
>
> *Machine Learning R&D level 1: Could significantly accelerate AI research at a cutting-edge lab if deployed widely, e.g. improving the pace of algorithmic progress by 3X, or comparably accelerate other AI research groups.*
>
> *The mismanagement of a model with these capabilities could enable the proliferation of cutting-edge AI*

[13] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, … Yuanzhi Li. Textbooks Are All You Need. Microsoft Research. June 20, 2023.
[14] AI at Meta. The Llama 3 Herd of Models. July 23, 2024.
[15] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, Mike Lewis. Self-Alignment with Instruction Backtranslation. ICLR 2024.
[16] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, … Jared Kaplan. Constitutional AI: Harmlessness from AI Feedback. December 15, 2022.
[17] Nvidia. Nemotron-4 340B Technical Report. August 6, 2024.
[18] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, Anima Anandkumar. Eureka: Human-Level Reward Design via Coding Large Language Models. April 30, 2024. ICLR 2024.
[19] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, Jirong Wen. A survey on large language model based autonomous agents. Frontiers of Computer Science. March 22, 2024.
[20] Qian Huang, Jian Vora, Percy Liang, Jure Leskovec. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. April 14, 2024.
[21] David Owen. Interviewing AI researchers on automation of AI R&D. August 27, 2024.

> *systems to malicious actors by enabling their AI development in turn. This could result in increased possibilities of harm from AI misuse, if AI models at that point were exhibiting capabilities like the ones described in other CCLs.*
>
> *Machine Learning R&D level 2: Could fully automate the AI R&D pipeline at a fraction of human labor costs, potentially enabling hyperbolic growth in AI capabilities.*
>
> *This could give any actor with adequate computational resources the ability to reach capabilities more powerful than those in the other CCLs listed in a short amount of time. The mismanagement of a model with these capabilities could result in the proliferation of increasingly and unprecedentedly powerful systems, resulting in significant possibilities of harm via misuse.*

> OpenAI's Preparedness Framework, page 11:
>
> *[Autonomy Risk Level - High]*
> *Model can execute open-ended, novel ML tasks on a production ML codebase that would constitute a significant step on the critical path to model self improvement (e.g., an OpenAI pull request that trains a new model with highly diverse datasets to improve codebase editing performance, using internal tooling and documentation).*
>
> *[Autonomy Risk Level - Critical]*
> *Model can profitably survive and replicate in the wild given minimal human instruction, i.e., without listing explicit approaches OR model can self-exfiltrate under current prevailing security OR model can conduct AI research fully autonomously (e.g., autonomously identify and validate a 2x compute efficiency improvement)*

Model Weight Security

Outlines of the levels of model weight security needed at different levels of AI capabilities. If malicious actors steal the weights of models with capabilities of concern, they could misuse those models to cause severe harm. Therefore, as models develop increasing capabilities of concern, progressively stronger information security measures are recommended to prevent theft and unintentional release of model weights. Security risks can come from insiders, or they can come from external adversaries of varying sophistication, from opportunistic actors to top-priority nation-state operations.[22,23]

> Google DeepMind's Frontier Safety Framework, pages 3–4:

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

The table below describes levels of security measures that may be applied to model weights to prevent their exfiltration. This is an important measure 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 bad actors) to any critical capabilities.

[0: Status quo]
Industry standard development and enterprise controls.

[1: Controlled access]
Access Control List hygiene. Non-forgeable lineage of models.

[2: Lockdown of unilateral access]
Significant restrictions of unilateral access of model weights.

[3: High-trust developer environments]
For developers with unilateral access to raw models, protection against exfiltration and account compromise.

[4: Advanced protection]
Model weights are generally not accessible to humans, even non-unilaterally. Hardened software platforms and confidential-compute hardware make it difficult even for well-funded adversaries to find and exploit vulnerabilities.

OpenAI's Preparedness Framework, page 20:

If we reach (or are forecasted to reach) at least "high" pre-mitigation risk in any of the considered categories we will ensure that our security is hardened in a way that is designed to prevent our mitigations and controls from being circumvented via exfiltration (by the time we hit "high" pre-mitigation risk). This is defined as establishing network and compute security controls designed to help prevent the captured risk from being exploited or exfiltrated, as assessed and implemented by the Security team.

This might require:

- *increasing compartmentalization, including immediately restricting access to a limited nameset of people, restricting access to critical know-how such as algorithmic secrets or model weights, and including a strict approval process for access during this period*
- *deploying only into restricted environments (i.e., ensuring the model is only available for inference in restricted environments) with strong technical controls that allow us to moderate the model's capabilities*
- *increasing the prioritization of information security controls.*

Anthropic's Responsible Scaling Policy, page 4:

[Current Safety Level]

> *Harden security against opportunistic attackers.*
>
> *[At the next Safety Level, AI Safety Level 3]*
> *Harden security such that non-state attackers are unlikely to be able to steal model weights and advanced threat actors (e.g. states) cannot steal them without significant expense.*
>
> *Implement internal compartmentalization for training techniques and model hyperparameters.*

Model Deployment Mitigations

Outlines of model-level and deployment mitigations against unintended usage needed at different levels of AI capabilities. Developers can train models to decline harmful requests[24] or employ additional techniques[25,26] including adversarial training[27,28] and output monitoring.[29] For models with greater levels of harmful capabilities, these safety measures may need to pass certain thresholds of robustness, including expert and automated red-

[24] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. April 12, 2022.
[25] Andy Zou, Long Phan, Justin Wang, Derek Duenas, Maxwell Lin, Maksym Andriushchenko, Rowan Wang, Zico Kolter, Matt Fredrikson, Dan Hendrycks. Improving Alignment and Robustness with Circuit Breakers. July 12, 2024.
[26] Rishub Tamirisa, Bhrugu Bharathi, Long Phan, Andy Zhou, Alice Gatti, Tarun Suresh, Maxwell Lin, Justin Wang, Rowan Wang, Ron Arel, Andy Zou, Dawn Song, Bo Li, Dan Hendrycks, Mantas Mazeika. Tamper-Resistant Safeguards for Open-Weight LLMs. August 1, 2024.
[27] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, Dan Hendrycks. HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. February 6, 2024.
[28] Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, Stephen Casper. Targeted Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs. July 22, 2024.
[29] AI at Meta. Llama Guard and Code Shield.

teaming.[30,31,32,33,34] Note that deployment mitigations are only effective as long as the model weights are securely within the possession of the developer.[35,36]

> OpenAI's Preparedness Framework, page 14:
>
> *Mitigations*
> *A central part of meeting our safety baselines is implementing mitigations to address various types of model risk. Our mitigation strategy will involve both containment measures, which help reduce risks related to possession of a frontier model, as well as deployment mitigations, which help reduce risks from active use of a frontier model. As a result, these mitigations might span increasing compartmentalization, restricting deployment to trusted users, implementing refusals, redacting training data, or alerting distribution partners.*
>
> 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 4:
>
> *[Current Safety Level]*
> *Follow current deployment best practices e.g. model cards, acceptable use policies, misuse escalation procedures, vulnerability reporting, harm refusal techniques, T&S tooling, and partner safety evaluation.*

[30] See also the Frontier Model Forum's issue brief on red-teaming.
[31] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. December 20, 2023.
[32] T. Ben Thompson, Michael Sklar. Fluent Student-Teacher Redteaming. July 24, 2024.
[33] See Planning red teaming for large language models (LLMs) and their applications for guidance on general red teaming. In addition to manual red teaming, automated red teaming frameworks such as PyRIT, developed by Microsoft's AI Red Team, can provide additional assurance.
[34] Nathaniel Li, Ziwen Han, Ian Steneker, Willow Primack, Riley Goodside, Hugh Zhang, Zifan Wang, Cristina Menghini, Summer Yue. LLM Defenses Are Not Robust to Multi-Turn Human Jailbreaks Yet. Scale AI. August 27, 2024.
[35] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, Jeff Alstott. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RAND. May 30, 2024.
[36] Peter Henderson, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal. Safety Risks from Customizing Foundation Models via Fine-Tuning. January 11, 2024.

> *[At the next Safety Level, AI Safety Level 3]*
> *Implement strong misuse prevention measures, including internal usage controls, automated detection, a vulnerability disclosure process, and maximum jailbreak response times.*
>
> *Each deployed modality (e.g. API, fine-tuning) must pass intensive expert red-teaming and evaluation measures for catastrophic risks.*

> Google DeepMind's Frontier Safety Framework, page 4:
>
> *This table below describes levels of deployment mitigations that may be applied to models and their descendants to manage access to and limit the expression of critical capabilities in deployment. Critical capabilities may have closely associated positive capabilities, misuse of critical capabilities may be more or less difficult to distinguish from beneficial uses, and the overall risks of misuse may differ by deployment context. Because of this, the mitigation options listed below are illustrative and will need to be tailored to different use cases and risks.*
>
> *[0: Status quo.]*
> *Safety finetuning of models and filters against general misuse and harmful model behavior.*
>
> *[1: Mitigations targeting the critical capability]*
> *Application, where appropriate, of the full suite of prevailing industry safeguards targeting the specific capability, including safety fine-tuning, misuse filtering and detection, and response protocols. Periodic red-teaming to assess the adequacy of mitigations.*
>
> *[2: Safety case with red team validation]*
> *A robustness target is set based on a safety case considering factors like the critical capability and deployment context. Afterwards, similar mitigations as Level 1 are applied, but deployment takes place only after the robustness of safeguards has been demonstrated to meet the target. Some protection against inappropriate internal access of the critical capability, such as automated monitoring and logging of large-scale internal deployments, Security Level 2.*
>
> *[3: Prevention of access]*
> *Mitigations that allow for high levels of confidence that capabilities cannot be accessed at all. Technical options for this level of deployment safety are currently an open research problem. Highly restricted and monitored internal use, alongside high security.*

Conditions for Halting Development and Deployment

Commitments to stop deploying and, if necessary, halt development of AI models if the AI capabilities of concern emerge before the improved mitigations can be put in place. Deploying models with concerning capabilities would be unsafe if broad user access enables catastrophic misuse. Training the model further would be unsafe if it is on track to develop

greater capabilities of concern and if the model weights are at risk of being stolen, due to insufficient security measures.

> Anthropic's Responsible Scaling Policy, page 2:
>
> *Complying with higher ASLs [a system of measures to mitigate risks] is not just a procedural matter, but may sometimes require research or technical breakthroughs to give affirmative evidence of a model's safety (which is generally not possible today), demonstrated inability to elicit catastrophic risks during red-teaming (as opposed to merely a commitment to perform red-teaming), and/or unusually stringent information security controls. Anthropic's commitment to follow the ASL scheme thus implies that we commit to pause the scaling and/or delay the deployment of new models whenever our scaling ability outstrips our ability to comply with the safety procedures for the corresponding ASL.*
>
> Page 10:
>
> *We will manage our plans and finances to support a pause in model training if one proves necessary, or an extended delay between training and deployment of more advanced models if that proves necessary. During such a pause, we would work to implement security or other measures required to support safe training and deployment, while also ensuring our partners have continued access to their present tier of models (which will have previously passed safety evaluations).*

> OpenAI's Preparedness Framework, page 21:
>
> *Only models with a post-mitigation score of "high" or below can be developed further. In other words, if we reach (or are forecasted to reach) "critical" pre-mitigation risk along any risk category, we commit to ensuring there are sufficient mitigations in place for that model (by the time we reach that risk level in our capability development, let alone deployment) for the overall post-mitigation risk to be back at most to "high" level. Note that this should not preclude safety-enhancing development. We would also focus our efforts as a company towards solving these safety challenges and only continue with capabilities-enhancing development if we can reasonably assure ourselves (via the operationalization processes) that it is safe to do so.*

> Google DeepMind's Frontier Safety Framework, page 2:
>
> *A model may reach evaluation thresholds before mitigations at appropriate levels are ready. If this happens, we would put on hold further deployment or development, or implement additional protocols (such as the implementation of more precise early warning evaluations for a given CCL) to ensure models will not reach CCLs without appropriate security mitigations, and that models with CCLs will not be deployed without appropriate deployment mitigations.*

Full Capability Elicitation during Evaluations

Intentions to perform evaluations in a way that elicits the full capabilities of the model. It is generally challenging to comprehensively explore a model's capabilities, even in a specific domain such as cyberoffense. Without dedicated efforts to elicit full capabilities,[37] evaluations may significantly underestimate the extent to which a model has capabilities of concern.[38] This is because model capabilities can be substantially improved through post-training enhancements such as fine-tuning, prompt engineering, or agent scaffolding.[39] For example, if the model is fine-tuned to refuse harmful requests, it is possible that the harmful capability could still be accessed through jailbreaking prompts discovered later. Model capabilities can additionally be improved through tool usage, such as the ability to execute code[40] or search the web.

Capability elicitation can involve a variety of techniques. One is to fine-tune the model to not refuse harmful requests, to reduce the chance that refusals lead to underestimation of capabilities. Another is to fine-tune the model for improved performance on relevant tasks. Evaluations that incorporate fine-tuning help to anticipate potential risks if model weights are stolen. Such evaluations are especially relevant if end users can fine-tune the model or if the developer improves fine-tuning later.[41] When evaluating AI agents, capabilities can be influenced by the quality of prompt engineering, tools afforded to the model, and usage of inference-time compute.[42]

> Google DeepMind's Frontier Safety Framework, page 6:
>
> *Capability elicitation: We are working to equip our evaluators with state of the art elicitation techniques, to ensure we are not underestimating the capability of our models.*

> Anthropic's Responsible Scaling Policy, page 12:

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

> *An inherent difficulty of an evaluations regime is that it is not currently possible to truly upper-bound the capabilities of generative models. However, it is important that we are evaluating models with close to our best capabilities elicitation techniques, to avoid underestimating the capabilities it would be possible for a malicious actor to elicit if the model were stolen.*
>
> - *False negatives due to harmlessness: While there are commercial and research incentives to develop maximally effective post-training techniques, certain evaluations may result in false negatives when used on commercial models. For example, harmlessness techniques may cause the model to refuse to assist with dangerous activities even when the underlying capability is present. Proper effort must be invested to avoid this type of false negative.*
> - *Mid-training evaluations: For significant scale-ups, it may be necessary to perform evaluations mid-training. Such models may have capability limitations due to various (potentially slow or expensive) fine-tuning stages having not yet occurred, or because performance may not scale linearly with compute in the midst of training. For now, we commit to perform mid-training fine-tuning and evaluations which, combined with the safety buffer described above, are intended to mitigate the risk of passing the defined ASL-3 threshold mid-training. We expect to update our procedures in the future as we better understand how to perform mid-training evaluations, for example by adjusting task difficulty to account for the limitations of a mid-training model. At high safety levels, we may transition to doing full fine-tuning even for mid-training evals in order to further mitigate risks of underestimating capabilities.*

> OpenAI's Preparedness Framework, page 13:
>
> *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.*

Timing and Frequency of Evaluations

Requirements about when evaluations must be performed, before deployment, during training, and after deployment. Evaluations are important throughout the model development lifecycle. Evaluations can inform whether it is safe to deploy a model externally or internally, based on its capabilities for harm and the robustness of safety measures. Merely

possessing a model can also pose risk if it has hazardous capabilities and the information security measures are insufficient to prevent theft of model weights.[43]

All three frameworks require evaluations both whenever the effective compute used to train their most capable models has increased significantly, and at regular intervals to catch the possibility that their models have improved via improved elicitation methods, fine-tuning, and so on. Effective compute takes into account both increased training compute and improved algorithmic efficiency.[44]

> OpenAI's Preparedness Framework, page 13:
>
> *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.*

> Google DeepMind's Frontier Safety Framework, page 2:
>
> *The capabilities of frontier models are tested periodically to check whether they are approaching a CCL. To do so, we will define a set of evaluations called "early warning evaluations," with a specific "pass" condition that flags when a CCL may be reached before the evaluations are run again.*
>
> *We are aiming to evaluate our models every 6x in effective compute and for every 3 months of fine-tuning progress. To account for the gap between rounds of evaluation, we will design early warning evaluations to give us an adequate safety buffer before a model reaches a CCL.*
>
> *Effective compute is a measure of the performance of a foundation model that uses scaling laws to integrate model size, dataset size, algorithmic progress, and compute into a single metric. While there is no direct relationship between effective compute and model size, a rough estimate suggests that a 6x increase in effective compute would correspond to approximately 2-2.5x increase in model size.*

> Anthropic's Responsible Scaling Policy, page 12:
>
> *During model training and fine-tuning, Anthropic will conduct an evaluation of its models for next-ASL capabilities both (1) after every 4x jump in effective compute, including if this occurs mid-training, and (2) every 3 months to monitor fine-tuning/tooling/etc improvements.*
>
> *We define effective compute as roughly the amount of compute it would have taken to train a model if no improvements to pretraining or fine-tuning techniques are included. This is operationalized by tracking the scaling of model capabilities (e.g. cross-entropy loss on a test set).*

[43] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, Jeff Alstott. Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models. RAND. May 30, 2024.
[44] See also the Frontier Model Forum's issue brief on measuring training compute.

Accountability

Exploratory intentions to seek accountability for capability evaluations and other practices. External oversight of a policy's implementation and the results of evaluations may be sought out as an independent check on the company, and to ensure that it is following through on the commitment it has made. This oversight may take the form of third-party red-teaming and evaluations execution, auditing policy compliance and mitigations implementation, or publicly releasing the results of evaluations and risk assessments.

> Anthropic's Responsible Scaling Policy, page 15:
>
> *External verification: Due to the large potential negative externalities of operating an ASL-4 [a future AI Safety Level] lab, verifiability of the above measures should be supported by external audits.*

> OpenAI's Preparedness Framework, page 25:
>
> *Accountability:*
>
> - *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.*
> - *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 Frontier Safety Framework, page 6:
>
> *Involving external authorities and experts: We are exploring internal policies around alerting relevant stakeholder bodies when, for example, evaluation thresholds are met, and in some cases mitigation plans as well as post-mitigation outcomes. We will also explore how to appropriately involve independent third parties in our risk assessment and mitigation processes.*

Updating Policies over Time

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

Google DeepMind's Frontier Safety Framework, page 6:

The Framework is exploratory and based on preliminary research. We expect it to evolve substantially as our understanding of the risks and benefits of frontier models improves, and we will publish substantive revisions as appropriate.

Anthropic's Responsible Scaling Policy, page 3:

This document will be periodically updated as we learn more, according to an "Update Process" described below. Updates will involve both defining higher ASL levels, and making course corrections to existing levels and safety measures as we learn more. We also welcome input on this document from other groups working on AI risk assessment and safety/security measures…

Follow an "Update Process" for this document, including approval by the board of directors…. Any updates will be noted and reflected in this document before they are implemented. The [company website].

We expect most updates to this process to be incremental, for example adding a new ASL level or slightly modifying the set of evaluations or security procedures as we learn more about model safety features or unexpected capabilities.

However, in a situation of extreme emergency, such as when a clearly bad actor (such as a rogue state) is scaling in so reckless a manner that it is likely to lead to lead to imminent global catastrophe if not stopped (and where AI itself is helpful in such defense), we could envisage a substantial loosening of these restrictions as an emergency response. Such action would only be taken in consultation with governmental authorities, and the compelling case for it would be presented publicly to the extent possible.

OpenAI's Preparedness Framework, page 7:

As mentioned, the empirical study of catastrophic risk from frontier Al models is nascent. Our current estimates of levels and thresholds for "medium" through "critical" risk are therefore speculative and will keep being refined as informed by future research. For this reason, we defer specific details on evaluations to the Scorecard section (and this section is intended to be updated frequently).

page 12:

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…, 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… to help us understand if there are any emerging or understudied threats that we need to track.*