AI Safety Framework for Pre-Frontier AI – Example Draft

This is a basic template of a draft policy to evaluate and mitigate severe AI risks. By starting with preliminary, low-cost evaluations, it is adapted for foundation model developers that are training models which are not yet at the forefront of AI capabilities and risk. Many of the details are illustrative or placeholders and may be customized for specific needs. This template serves as an example to demonstrate one possible approach to managing severe AI risks. See also: Common Elements of Frontier AI Safety Policies.

Introduction

As our foundation models become more advanced over time, we are monitoring their potential to pose catastrophic risks[1] if, in the future, they have hazardous capabilities such as:[2]

- **Chem-bio weapons assistance**: The ability to enable a nontechnical expert to develop, acquire, or deploy a chemical or biological weapon; or to uplift experts to develop a top-severity chemical or biological weapon.

[1] Definitions of catastrophic risk vary, but one definition might be risks that could cause thousands of deaths or the equivalent of billions of dollars of damage.
[2] Compare to the risks discussed in the G7 Hiroshima Process, which various advanced AI organizations commit to identifying and mitigating.

- **Advanced cyberoffense**: The ability to uplift cyberoffense capabilities to disable critical infrastructure at scale, discover high-value vulnerabilities, and so on.
- **Automated AI R&D**: The ability to automate the development and capabilities advancement of frontier AI (for example, including AI with hazardous capabilities), in a way that could outpace safety and alignment measures.
- [additional hazardous capabilities may be listed]

Our policy here intends to manage these potential risks from our models through commitments to:

- Evaluate the capabilities of our models, including hazardous capabilities, during development, prior to deployment, and after deployment.
- Adopt robust safety and security mitigations before our models develop hazardous capabilities.
    - The safety mitigations would help protect against catastrophic misuse or certain autonomous AI risks.
    - The security mitigations would prevent sophisticated adversaries from stealing our model weights and misusing them.
- Provide transparency about our hazardous capability evaluations and mitigations.

Our policy sets forth two sets of measures:

- Pre-Frontier Measures, which take effect immediately. These include low-cost evaluations and are suitable for when our models are substantially behind the capabilities of cutting-edge AI models or any hazardous AI models.
- Frontier Measures, which take effect when our models are no longer pre-frontier. These include more thorough evaluations, as safety and security mitigations depending on specific dangerous capabilities.

Under our Pre-Frontier Measures, we believe that our model is safe to develop and deploy (at least with respect to catastrophic risk) if all of the following hold:

- Our models are below our General Capability Threshold, which indicates they are behind the frontier of general AI capabilities of early 2024 [for example].
- Our models are below our Pre-Hazard Thresholds, scoring substantially lower than frontier models of early 2024 and human experts on low-cost benchmarks for capabilities of concern.
- Our models are not specially trained on chemical or biological data (e.g., biological sequence data).[3]

[3] Biological design tools may need additional evaluation and mitigation measures beyond what is discussed in this template. Related: Developing Guardrails for AI Biodesign Tools.

Once any of these conditions fail to hold, then our Frontier Measures take effect. As part of our Frontier Measures, we will conduct more intensive evaluations to check whether our models have hazardous capabilities (approaching our Hazardous Capability Thresholds) and adopt corresponding safety and security mitigations.

Update Process

The science of capability evaluations and threat assessment is emerging, and we plan to update our practices frequently.

Any updates to our practices will be reviewed and approved by [committee] at least two weeks before the updates are adopted. Updates will be made public at the same time that we adopt them, and linked here. We will maintain a changelog and links to previous versions.

Pre-Frontier Measures

In our Pre-Frontier Measures, which are active immediately, we conduct ongoing pre-deployment evaluations during development and prior to deployment to check that our most advanced models are below our General Capability Threshold and Pre-Hazard Thresholds. We also check that they are not specially trained on biological data (more so than a typical large language model). Otherwise, our Frontier Measures enter into force.

The scores in our General Capability Threshold and Pre-Hazard Thresholds are set conservatively to err on the side of safety. We recognize that simple benchmarks are only an approximate proxy for risky capabilities. Our evaluations may use chain-of-thought, consensus@k, etc. where appropriate.

As part of our evaluations, we will develop scaling laws to forecast how much training compute our models would need to exceed these thresholds.

[If agent benchmarks such as GAIA or Cybench are included] In our agent evaluations, we also engage in capability elicitation efforts to avoid underestimating model capabilities, by experimenting with multiple agent scaffolds, training models for agentic tool usage, ensuring that the models have high-quality agent scaffolds, reviewing transcripts to check for simple failures, and so on.

General Capability Threshold

To be below the General Capability Threshold, a model must underperform all of these benchmark scores in the Score column which are conservatively below the state-of-the-art of early 2024.

[The benchmarks listed here are just examples.]

Category	**Score**	**Reference scores**[4, 5]
Question-answering	MMLU: _____ SimpleQA: _____	MMLU: 91.8% by o1, 85.7% by gpt-4o-2024-05-13, 88.3% by Claude 3.5 Sonnet SimpleQA: 42.6% by o1, 7.6% by o1-mini, 39.0% by gpt-4o-2024-05-13, 28.9% by Claude 3.5 Sonnet
Math	MATH: _____	76.6% by gpt-4o-2024-05-13 76.1% by Grok 2
Programming	HumanEval: _____	91.0% by gpt-4o-2024-05-13
[and so on]		

Pre-Hazard Thresholds

A model must underperform on all benchmark scores in the Pre-Hazard Threshold column to be considered below the Pre-Hazard Thresholds. As a conservative measure, we set our Pre-Hazard Thresholds significantly below human expert performance and reference models.

Category	**Pre-Hazard Threshold**	**Reference scores**
CBRN weapons	LAB-Bench ProtocolQA (multiple-choice): _____ BioLP-Bench: _____	LAB-Bench ProtocolQA: 79% by human expert, 73.1% by o1-preview (post-mitigation, consensus@32), 57.4% by GPT-4o (consensus@32)

[4] See here for a table of reported benchmark scores for various models.
[5] Artificial Analysis

		BioLP-Bench: 38.4% by average expert, 33% by o1 (post-mitigation, pass@1), 20% by GPT-4o (pass@1)
Cyberoffense	eyeballvul: _____ [or other cyber benchmarks, such as Cybench]	eyeballvul: F1 scores of 14.1% by Claude 3 Opus, 13.1% by Claude 3.5 Sonnet (original)
AI R&D	LiveCodeBench: _____ [or other programming or machine learning engineering benchmarks, such as RE-Bench or MLE-bench Lite]	LiveCodeBench: 48.8% by Claude 3.5 Sonnet (original, pass@1, 04/01/2024 – 06/01/2024)

Operations

[role/team] will be responsible for ensuring that we execute these evaluations. We will make our model scores on these benchmarks public along with any model releases. We may update our threshold definition to integrate new benchmarks or modify scores, following our Update Process.

We will consult with [third-party auditor] to provide information on our model scores on the above benchmarks and our scaling laws or estimates of when the General Capability Threshold or Pre-Hazard Thresholds may be surpassed, every [3–6 months].

If our models reach the General Capability Threshold or Pre-Hazard Thresholds, [role/team] will notify the CEO, [third-party auditor], and the broader organization that the Frontier Measures will take effect immediately. Prior to that point, [role/team] will share with the CEO and organization periodic forecasts of when the General Capability Threshold or Pre-Hazard Thresholds may be reached, or if they may be imminently reached, in order to prepare the organization.

Frontier Measures

[Note: The focus of this template is more on the Pre-Frontier Measures, while the Frontier Measures described here may not be adequate for AI labs developing the most advanced models. We suggest considering ways to make these measures considerably more detailed with

stronger evaluation, transparency, and oversight processes, whether in an initial or updated policy.]

Hazardous Capability Thresholds

The table below lists our Hazardous Capability Thresholds [with fill-in-the-blank definitions]. These thresholds are associated with security requirements, which are to be achieved before we develop a model that reaches the threshold. They also have requirements for model safety which are required before internal or external deployment. The example evaluations listed can be substituted with easier tasks that assess for capabilities prerequisite to the hazardous capability.

Hazardous Capability Threshold[6]	**Security Requirements**[7]	**Internal and External Deployment Requirements**	**Example Evaluations**
Chem-bio weapons assistance: _____	Security to prevent theft of model weights by most cybercrime organizations or insider threats.	Before external deployment, safeguards (such as model refusals or input/output filters) are robust enough to avoid meaningful uplift of malicious actors that enables them to create or deploy a chem-bio weapon.	Assign technical non-experts to perform safe tasks representative of chem-bio weaponization skills, comparing performance with and without model assistance. Evaluate models on question-answering or lab automation tasks, or their ability to improve human performance on such tasks.
Advanced cyberoffense: _____	Before training more capable models or broad external deployment, study whether the model could lead to catastrophic impacts in		Evaluate models on cyber offense tasks such as vulnerability

[6] See A Sketch of Potential Tripwire Capabilities for AI for example threshold definitions.
[7] Consider the security levels discussed in Securing AI Model Weights, such as Security Level 3 to protect against most malicious actors and insider threats, Security Level 4 to protect against standard operations by leading cyber organizations, or Security Level 5 to protect against top-priority operations by leading cyber organizations.

	cybersecurity, disproportionate to its potential to improve cybersecurity or to society's existing cyberdefenses. If so, consider adopting requirements for model weight security, model safety, or limited deployment to verified users.		detection or exploitation, up to the complexity of what leading experts or cyber-capable organizations can complete.
Automated AI R&D: _____	Security to prevent theft of model weights by leading cyber-capable organizations, including top-priority operations of nation-state cyber groups.	Before deploying the model internally, publish a plan to ensure that alignment techniques keep pace with advancements in AI capabilities. Internal deployment also requires a safety case[8, 9] (based on inability, control, or alignment) that the model will not cause unacceptable outcomes such as sabotaging safety research or triggering unauthorized deployments without safety measures.	Evaluate models on frontier machine learning engineering and research tasks.

[Descriptions of how these thresholds were determined]

[Example scenarios where the models or systems would pose intolerable risk]

Early warning protocol

We will:

- Perform capability evaluations to test whether our models might have these capabilities, in cases where there is not already strong reason to think they lack these capabilities.
- Require pre-scaleup safety cases before further training our leading models. Prior to scaling up a leading model, we will write a safety case, which incorporates evidence from evaluation results, scaling laws, and model transcripts, which establishes that our next round of scaleup will not surpass any of our Hazardous Capability Thresholds

[8] Safety Cases: How to Justify the Safety of Advanced AI Systems

[9] Towards evaluations-based safety cases for AI scheming

without our adopting the necessary mitigations. The pre-scaleup safety case would need to be reviewed and approved by a third party.
- Keep risk mitigations in line with capabilities: when our models do have these capabilities (or when we cannot establish reasonably strong evidence that they do not[10]), we will implement the needed mitigations, halting development and deployment of frontier AI models as needed until we can do so.

The capability evaluation process (CEP)

As part of our capability evaluation process (CEP), we will run evaluations, possibly partnering with other organizations, to determine whether an AI model could meet any of our Hazardous Capability Thresholds. This process will consist of:

- Determining whether there are tasks that can provide *strong evidence that a model lacks the required capabilities*. These would be tasks that (a) seem prerequisite to the capabilities we're testing for (in the sense that an AI that could not do the task very likely lacks the capability); (b) are *robustly* not doable by the AI model in question.
    - We provide illustrative examples of these tasks in [appendix].
    - In order to establish that a task is robustly not doable, we will need to show that it is not doable even with a significant effort to get the AI to do the task, of the kind that a malicious actor with access to the full model weights and code might attempt. This could include fine-tuning the model to attempt even harmful tasks; ensuring that the model is prompted and scaffolded in whatever way will lead to the best performance, including e.g. a multi-agent setup; connecting the model to whatever tools and plugins will maximize its odds of success; etc.
    - We will sometimes be able to make a case that a model lacks a capability of interest by comparing its *general* capabilities to those of another model, for which evaluations have already been run and reported on.
- Developing scaling laws of how performance on hazardous capability assessments vary with increased training compute or capability elicitation effort.
- Writing up a report on what tasks the model robustly cannot complete, and which capabilities it likely lacks. The report will include a pre-scaleup safety case with a justification of how much scaleup we believe would be safe without exceeding Hazardous Capability Thresholds that we have not already implemented mitigations for.
- Submitting the report to our [committee] and [third-party organization], which will review it, ask for more information as warranted, and make a final determination about which capabilities can be ruled out for the model.

[10] Including with respect to a safety margin, discussed more below.

To the extent that key capabilities cannot be ruled out, we will either implement the risk mitigations they require or halt development and deployment until we can do so, as specified in the next section.

Capability reports, redacted as appropriate to protect sensitive information and the public interest, will be published within 2 weeks of when both (a) the reports have been internally approved (b) the model in question is commercially deployed.

We will perform our capability evaluations process (CEP) for the first model that exceeds our General Capability Thresholds or Pre-Hazard Thresholds, as well as any model that is trained with at least 4× more effective compute than the last model that was evaluated with the CEP. We will also perform the CEP at least once every 6 months, to monitor the impact of improved post-training enhancements.[11]

Halting development and deployment

We will:

- Halt frontier AI development (improvement of the model's general capabilities, and of other models of comparable or greater capabilities) of a model further than allowed by a pre-scaleup safety case that has been approved by [committee] and [third-party organization].
- Not deploy a model until we have confirmed that the security level is in line with its capabilities.
- Halt further AI frontier development if we cannot make such a confirmation of the security level within 4 weeks of beginning the CEP. The halt in development will last until we can make such a confirmation.

Accountability

This document has been made publicly available, and relevant employees have been given a clear procedure for reporting any practices they believe are inconsistent with it, including anonymously if they choose to do so.

A specific person has been named as having primary responsibility for ensuring that:

[11] Post-training enhancements are discussed in AI capabilities can be significantly improved without expensive retraining.

- The capability evaluations process is run on any models fitting the criteria described in the previous section, both prior to initial deployment and every 6 months afterward.
- The capability evaluations process involves a good-faith attempt to determine what capabilities our models could have if a serious effort were made to elicit these capabilities from them.
- Risk mitigations are kept in line with capabilities, including via restricting deployment and development as specified by the policy, and are sufficient to accomplish the goals laid out in [Security Levels appendix].

We will engage an external organization to audit our compliance with this policy yearly. When we intend to have upgraded our security level, we will also engage a security organization to audit our security level.

To promote greater transparency on how we interpret our Hazardous Capability Thresholds and our progress towards meeting future safety and security requirements, we publish:

- *Early warning thresholds*: Quantitative scores on public and private benchmarks such that, if the model scores below these thresholds, we feel confident it has not reached the Hazardous Capability Threshold. For private benchmarks still in development, early warning thresholds can be defined by human expert performance or degree of uplift to human performance.
- *Hazardous capability indicators*: Quantitative benchmark scores and other indicators that reflect our best estimate of a model's performance if it met our Hazardous Capability Thresholds.
- *Security roadmap*: A projected timeline for achieving specific security levels in line with expected advancements in model capabilities.
- *Safety KPIs*: Target metrics and indicators measuring progress in areas such as safeguard robustness, jailbreak response time, detection and mitigation of AI scheming, and so on.
- *Capability and alignment profile of AI R&D agents*: Description of the capabilities and alignment properties of our models capable of automated AI R&D, operationalized as models (including nonpublic models) that achieve above x% on [machine learning engineering benchmark].

These will be described in an up-to-date document at [link], with changes recorded in a changelog and justified publicly.