Frontier Safety Framework

Version 1.0

The Frontier Safety Framework is our first version of a set of protocols that aims to address severe risks that may arise from powerful capabilities of future foundation models. In focusing on these risks at the model level, it is intended to complement Google's existing suite of AI responsibility and safety practices, and enable AI innovation and deployment consistent with our AI Principles.

In the Framework, we specify protocols for the detection of capability levels at which models may pose severe risks (which we call "Critical Capability Levels (CCLs)"), and articulate a spectrum of mitigation options to address such risks. We are starting with an initial set of CCLs in the domains of Autonomy, Biosecurity, Cybersecurity, and Machine Learning R&D. Risk assessment in these domains will necessarily involve evaluating cross-cutting capabilities such as agency, tool use, and scientific understanding. We will be expanding our set of CCLs over time as we gain experience and insights on the projected capabilities of future frontier models.

We aim to have this initial framework implemented by early 2025, which we anticipate should be well before these risks materialize. The Framework is exploratory and based on preliminary research, which we hope will contribute to and benefit from the broader scientific conversation. It will be reviewed periodically and we expect it to evolve substantially as our understanding of the risks and benefits of frontier models improves.

The Framework is informed by the broader conversation on Responsible Capability Scaling.[1] The core components of Responsible Capability Scaling are to:

- Identify capability levels at which AI models pose heightened risk without additional mitigations
- Implement protocols to detect the attainment of such capability levels
- Prepare and articulate mitigation plans in advance for when such capability levels are attained
- Where appropriate, involve external parties in the process to help inform and guide our approach.

We are piloting this initial version of the Frontier Safety Framework as a first step to operationalizing these principles.

Table of contents:

Framework **2**

Mitigations **3**

Security Mitigations 3

Deployment Mitigations 4

Critical Capability Levels **4**

Future work **6**

Acknowledgements **7**

[1] See https://www.gov.uk/government/publications/emerging-processes-for-frontier-ai-safety, https://metr.org/blog/2023-09-26-rsp/, https://www.anthropic.com/news/anthropics-responsible-scaling-policy, https://openai.com/preparedness/

Framework

This section describes the central components of the Frontier Safety Framework.

1 - Critical Capability Levels:
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,[2] 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. The CCLs we have identified are described below.

2 - Evaluating frontier models:
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[3] 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.[4]

3 - Applying mitigations:
When a model reaches evaluation thresholds (i.e. passes a set of early warning evaluations), we will formulate a response plan based on the analysis of the CCL and evaluation results. We will also take into account considerations such as additional risks flagged by the review and the deployment context.

The initial version of the Framework focuses on two categories of mitigations: *security mitigations* to prevent the exfiltration of model weights, and *deployment mitigations* (such as safety fine-tuning and misuse filtering, detection, and response) to manage access to and prevent the expression of critical capabilities in deployments. We have developed frameworks for Security Levels and Deployment Levels to enable calibrating the robustness of mitigations to different CCLs.

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.

Figure 1 depicts the relationship between these components of the Framework.

---

[2] "Autonomy" captures the potential misuse of AI models with significant capacity for agency and flexible action/tool use over long time horizons and across multiple domains.
[3] 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.
[4] More specifically, the early warning evaluations will be a set of model evaluations that we are confident will be passed before the model is 6x in effective compute or 3 months of fine-tuning away from the CCL.

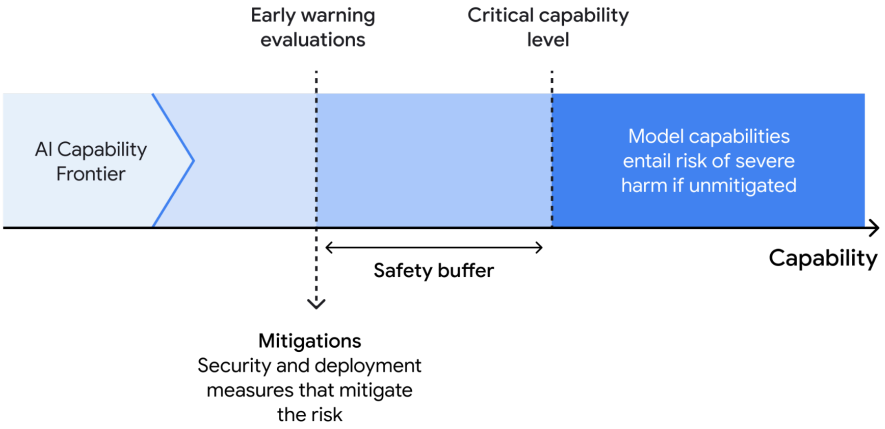

Figure 1: the relationship between different components of the Framework.

Mitigations

The Frontier Safety Framework proposes two kinds of mitigations to address models with critical capabilities: *security mitigations* to prevent the exfiltration of model weights, and *deployment mitigations* to manage access to/prevent the expression of critical capabilities in deployments. For each category of mitigation, we have developed several levels of mitigations, allowing us to calibrate the robustness of measures to the risks posed.

Security Mitigations

The table below describes levels of security mitigations 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.

Level and capabilities	Measures
0: Status quo	Industry standard development and enterprise controls. E.g., multi-factor authentication, basic access control mechanisms, secure software development standards, red-team tests.
1: Controlled access Access Control List hygiene. Non-forgeable lineage of models. Approximately RAND L3.[5]	Limited access to raw representations of the most valuable models, including isolation of development models from production models. Specific measures include model and checkpoint storage lockdown, SLSA Build L3 for model provenance, and hardening of ML platforms and tools.
2: Lockdown of unilateral access Significant restrictions of unilateral access of model weights. Approximately RAND L3-L4.	Changes to ML platforms and tools to disallow unilateral access to raw model representations outside the core research team, with exceptions granted on the basis of business need.
3: High-trust developer environments For developers with unilateral access to raw models, protection against exfiltration and account compromise. Approximately RAND L4.	Models can be accessed only through high-trust developer environments (HTDE), hardened, tamper-resistant workstations with enhanced logging.

[5] See https://www.rand.org/pubs/working_papers/WRA2849-1.html.

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. Approximately RAND L5.	Minimal trusted computing base (TCB). TPUs with confidential compute capabilities. Dedicated hardware pods for training and serving high-value models.

Deployment Mitigations

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.

Level and capabilities	Measures
0: Status quo	Safety finetuning of models and filters against general misuse and harmful model behavior.
1: Mitigations targeting the critical capability Use of the full suite of mitigations to prevent the inappropriate access of critical capabilities.	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 Targeted safeguards, aimed at keeping numbers of incidents below a prespecified amount, with pre-deployment validation by a red-team.	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.

Critical Capability Levels

Critical Capability Levels describe thresholds at which models may pose heightened risk without additional mitigation.[6] We will develop early warning evaluations to detect when models approach CCLs, and apply appropriate mitigations to models that reach evaluation thresholds.

[6] Note: when we refer to a model's capabilities, we include capabilities resulting from any reasonably foreseeable fine-tuning and scaffolding to turn the model into a functioning system.

The table below details an initial set of CCLs we have identified through a preliminary analysis of the Autonomy, Biosecurity, Cybersecurity, and Machine Learning R&D risk domains. As we conduct further research into these and other risk domains, we expect these CCLs to evolve and for several CCLs at higher levels or in other risk domains to be added.

Risk domain	Critical capability level	Rationale
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.
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.
Cybersecurity: Risks of models assisting in the execution of a cyber attack.	**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.
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	**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	The mismanagement of a model with these capabilities could enable the proliferation of cutting-edge AI 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

attainment or proliferation of other powerful AI models.	algorithmic progress by 3X, or comparably accelerate other AI research groups.	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.

Future work

We aim to have this initial framework implemented by early 2025, which we anticipate should be well before these risks materialize.

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. Issues that we aim to address in future versions of the Framework include:

- **Greater precision in risk modeling:** Given the nascency of the underlying science, there is significant room for improvement in understanding the risks posed by models in different domains, and refining our set of CCLs. We also intend to take steps to forecast the arrival of CCLs to inform our preparations.
- **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.
- **Mitigation plans:** Striking a balance between mitigating risks and fostering access and innovation is crucial, and requires consideration of factors like the context of model development, deployment, and productization. As we better understand the risks posed by models at different CCLs, and the contexts in which our models will be deployed, we will develop mitigation plans that map the CCLs to the security and deployment levels described.
- **Updated set of risks and mitigations:** There may be additional risk domains and critical capabilities that fall into scope as AI capabilities improve and the external environment changes. Future work will aim to include additional pressing risks, with possible examples including:
    - **Misaligned AI:** protection against the risk of systems acting adversarially against humans may require additional Framework components, including new evaluations and control mitigations that protect against adversarial AI activity.
    - **Chemical, radiological, and nuclear risks:** Powerful capabilities in each of these domains are currently covered by existing model evaluations that Google DeepMind is already implementing, and they are potential candidates for inclusion into the Framework.
    - **Higher CCLs:** as AI progress continues, we may approach more advanced capabilities within existing domains, especially when present CCLs are close to being breached.
- **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.

Acknowledgements

The Frontier Safety Framework was developed by Lewis Ho, Rohin Shah, Celine Smith, Seb Farquhar, Seb Krier, Dave Orr, Max Poletto, and Claudia van der Salm, under the leadership of Allan Dafoe, Helen King, Tom Lue, Nicklas Lundblad, Demis Hassabis, Anca Dragan, and Shane Legg. We are grateful for the contributions of Matthew Aitchison, Joslyn Barnhart, Jennifer Beroshi, Jarrah Bloomfield, Dawn Bloxwich, Vijay Bolina, Sasha Brown, Karen Courington, Seliem El-Sayed, Iason Gabriel, Conor Griffin, Will Hawkins, Heidi Howard, Alex Kaskasoli, Christoph Kern, Lucy Lim, Philip Mackenzie, James Manyika, Aalok Mehta, Sam Pettit, Trung Phan, Mary Phuong, Toby Shevlane, Alexandru Totolici, Kent Walker, and Don Wallace, with input from many others. We would like to thank METR for contributing their expertise on Responsible Capability Scaling.