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:

[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/

4th February 2025

Frontier Safety Framework

Version 2.0

The Frontier Safety Framework is a set of protocols that aims to address severe risks that may arise from powerful capabilities of foundation models. 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.

The Framework is informed by the broader conversation on Frontier AI Safety Frameworks.[1] 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.

In version 2.0 of 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 mitigation approaches to address such risks. At present, the Framework primarily addresses misuse risk,[2] but we also include an exploratory section addressing deceptive alignment risk,[3] focusing on capability levels at which such risks may begin to arise. For each type of risk, we define here a set of CCLs and a mitigation approach for them. Risk assessment will necessarily involve evaluating cross-cutting capabilities such as agency, tool use, reasoning, and scientific understanding.

The safety of frontier AI systems is a global public good. The protocols here represent our current understanding and recommended approach of how severe frontier AI risks may be anticipated and addressed. Importantly, there are certain mitigations whose social value is significantly reduced if not broadly applied to AI systems reaching critical capabilities. These mitigations should be understood as recommendations for the industry collectively: our adoption of them would only result in effective risk mitigation for society if all relevant organizations provide similar levels of protection, and our adoption of the protocols described in this Framework may depend on whether such organizations across the field adopt similar protocols.

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.

[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/, https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/.

[2] As in, in the context of the Framework, risks of threat actors using critical capabilities of deployed or exfiltrated models to cause harm.

[3] As in, in the context of the Framework, risks of highly autonomous systems purposefully undermining human control over AI systems.

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.

Table of Contents:

Framework

This section describes the central components of the Frontier Safety Framework. These protocols represent our current understanding of and recommended approach for how severe frontier AI risks may be anticipated and addressed.

1 - Critical Capability Levels
The Framework is built around capability thresholds called "Critical Capability Levels." These are capability levels at which, absent mitigation measures, AI models or systems may pose heightened risk of severe harm. CCLs can be determined by identifying and analyzing the main foreseeable paths through which a model could cause severe harm, and then defining the CCLs as the minimal set of capabilities a model must possess to do so. Note that we have selected our CCLs to be conservative; it is not clear to what extent CCLs might translate to harm in real-world contexts.

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

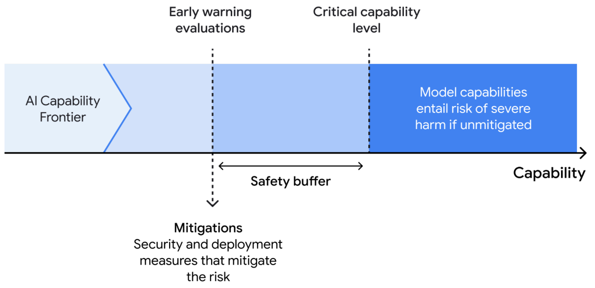
Early warning evaluations
Critical capability level
AI Capability Frontier
Model capabilities entail risk of severe harm if unmitigated
Safety buffer
Capability
Mitigations
Security and deployment measures that mitigate the risk

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.

delineating when such capability becomes present, and subsequently when the initial mitigation for this capability—automated monitoring—is no longer adequate.

2 - Assessing the Capabilities of Frontier Models

We intend to evaluate our most powerful frontier models regularly to check whether their AI capabilities are approaching a CCL. We also intend to evaluate any of these 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 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. We may run early warning evaluations more frequently or adjust the alert threshold of our evaluations if the rate of progress suggests our safety buffer is no longer adequate.

Where necessary, early warning evaluations may be supplemented by other evaluations to better understand model capabilities relative to our CCLs. We may use additional external evaluators to test a model for relevant capabilities, if evaluators with relevant expertise are needed to provide an additional signal about a model's proximity to CCLs.

3 - Applying Mitigations

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. This will inform the formulation and application of a response plan.

Central to most response plans will be the application of the mitigations described later in this document. For misuse, we have two categories of mitigations: *security mitigations* intended to prevent the exfiltration of model weights, and *deployment mitigations* (such as safety fine-tuning and misuse filtering, detection, and response) intended to counter the misuse of critical capabilities in deployments. For deceptive alignment risk, *automated monitoring* may be applied to detect and respond to deceptive behavior for models that meet the first deceptive alignment CCL. Note that these mitigations reflect considerations from the perspective of addressing severe risks from powerful capabilities alone; due to this focused scope, other risk management and security considerations may result in more stringent mitigations applied to a model than specified by the Framework.

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. Conversely, where model capabilities remain quite distant from a CCL, a response plan may involve the adoption of additional capability assessment processes to flag when heightened mitigations may be required.

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.

Misuse

This section describes our mitigation approach for models that pose risks of severe harm through misuse, and then details our set of misuse CCLs, as well as the mitigations that we provisionally assess as appropriate for them.

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.

Mitigation Approach

There are two categories of mitigations to address models with misuse critical capabilities: *security mitigations* intended to prevent the exfiltration of model weights, and *deployment mitigations* intended to counter the misuse of critical capabilities in deployments. For security, we have several levels of mitigations, allowing calibration of the robustness of security measures to the risks posed. For deployment mitigations, we specify a standard process for applying, assessing, and reviewing mitigations: the aim of this process is to calibrate mitigations to CCLs, and the procedural approach reflects the more iterative and CCL-dependent nature of deployment mitigations.

Security Mitigations

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[4] 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.[5] Because AI security is an area of active research, we expect the concrete measures implemented to reach each level of security to evolve substantially.

Deployment Mitigations

The following deployment mitigation process will be applied to models reaching a CCL, allowing for iterative and flexible tailoring of mitigations to each risk and use case.[6]

1. **Development and assessment of mitigations:** safeguards and an accompanying safety case[7] 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[8] 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.

Misuse Critical Capability Levels

The table below details a set of CCLs we have identified through ongoing analysis of the CBRN, Cyber, and Machine Learning R&D risk domains. We expect these to evolve over time. We recommend a security level to each of these CCLs, which reflect our assessment of the minimum appropriate level of security

[4] See https://www.rand.org/pubs/research_reports/RRA2849-1.html, pp 21-22.

[5] As the authors point out, the "security level benchmarks represent neither a complete standard nor a compliance regime—they are provided for informational purposes only and should inform security teams' decisions rather than supersede them."

[6] This process is derived from that required by Deployment Level 2 in Frontier Safety Framework version 1.0. Our rationale for this change is as follows: the process extends naturally to cover risks for which "prevention of access" (Deployment Level 3) is required, and is worth applying to all CCLs.

[7] A safety case is an assessable argument showing how severe risks associated with a model's CCLs have been minimised to an appropriate level. See also, for reference, https://www.aisi.gov.uk/work/safety-case-template-for-inability-arguments and https://arxiv.org/abs/2410.21572.

[8] General availability deployment refers to public deployments intended for production use and without the imposition of significant user or use restrictions.

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

the field of frontier AI should apply to models reaching each CCL. In practice, our overall security posture may commonly exceed the baseline levels recommended here.

These recommended security levels reflect our current thinking and may be adjusted if our empirical understanding of the risks changes. This may occur if, for example, a model does not possess capabilities meaningfully different from other widely available models that have not demonstrably caused or contributed to severe risks, or if we assess that the benefits of the open release of model weights outweigh the risks. Relatedly, we believe these recommendations will only be effective if the entire frontier AI field applies them, and of limited social utility if not.

Table 1: Misuse CCLs and Security Mitigations

Risk domain[9]	Critical capability level	Recommended security level and rationale
CBRN: Risks of models assisting in the development, preparation and/or execution of CBRN attacks.	**CBRN uplift 1:** 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.[10]	**Security controls and detections at a level generally aligned with RAND SL 2** The potential magnitude of harm these capabilities may enable means the exfiltration and leak of model weights reaching this CCL could be highly damaging. However, low-resourced actors are unlikely to pose a substantial exfiltration threat.
Cyber: Risks of models assisting in the execution of a cyber attack.	**Cyber autonomy level 1:** Can be used to drastically reduce the cost (e.g. through full automation) of carrying out end-to-end cyberattacks[11] on organizations with a limited security posture.	**Security controls and detections at a level generally aligned with RAND SL 2** 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. Exfiltration of model weights could enable the execution of such attacks at scale. However, cybersecurity may improve correspondingly when models reach such capability levels. The relatively ambiguous net costs of exfiltration count against security levels with higher costs to innovation.
	Cyber uplift level 1: Can be used to significantly assist with high impact cyber attacks,[12] resulting in overall cost/resource reductions of an order of magnitude or more.[13]	**Security controls and detections at a level generally aligned with RAND SL 2** A model at this capability level could help fairly well-resourced threat actors carry out severe cyber attacks on targets like critical businesses, national government entities, and critical national infrastructure with lower resource expenditure, potentially increasing the frequency of such attacks significantly. However, as above, cyber defense may improve to diminish the impact of

[9] Note that we have removed the Autonomy risk domain, which was included in Frontier Safety Framework version 1.0. Most of the advanced risk that was captured by this CCL is now covered by our misalignment section. From the perspective of misuse risks, our threat models suggest that no heightened deployment mitigations would be necessary, and that security controls and detection at a level generally aligned with RAND SL 2 would be adequate.

[10] For example, through the use of a self-replicating CBRN agent. Compared to a counterfactual of not using generative AI systems.

[11] E.g. deletion or exfiltration of critical information, or destroying or disabling key systems.

[12] E.g. deletion or exfiltration of sensitive information/disruption of key systems of organizations with strong security posture, or the creation of predictable and large-scale effects on critical national infrastructure.

[13] Relative to the counterfactual of using 2024 AI technology and tooling.

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.

		AI-assisted cyber attacks. Similarly, the ambiguous net costs of exfiltration count against security levels with higher costs to innovation.
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 uplift level 1:** Can or has been used to accelerate AI development, resulting in AI progress substantially accelerating (e.g. 2x) from 2020-2024 rates.	**Security controls and detections at a level generally aligned with RAND SL 3** Unrestricted access to models at this level of capability could significantly increase an actor's ability to progress to yet more powerful models and other critical capabilities. The exfiltration of such a model may therefore have a significant effect on society's ability to adapt to and govern powerful AI systems, effects that may have long-lasting consequences. Substantially strengthened security is therefore recommended.
	Machine Learning R&D autonomy level 1: Can fully automate the AI R&D pipeline at a competitive cost. The cost comparison is relative to humans augmented by AI tools.	**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.** Unrestricted access to models at this level of capability could give an actor with adequate computational resources the ability to reach capabilities much more powerful than those in the other CCLs listed in a short amount of time. This could be catastrophic if there is no effective way of defending against rapidly improving and potentially superhuman AI systems wielded by threat actors. Therefore, we recommend models at this level of capability have exceptional security even though they may have substantial innovation costs.

Deceptive Alignment

This section describes an initial approach for addressing risks of deceptive alignment. By "deceptive alignment," we mean the risk that AI systems purposefully undermine human control over AI systems. While we do not express any opinion here about how likely it is for deceptive alignment to arise, this section's provisions aim to prepare for this possibility.

Note that the effective mitigation of deceptive alignment risk requires the application of safeguards to all AI systems reaching critical capabilities. It is important that the industry adopts strategies for addressing this risk: what is detailed in this section represents one possible approach for doing so.

Mitigation Approach and Critical Capability Levels

An initial mitigation 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. 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.

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.

The table below summarizes the initial approach:

Table 2: Deceptive Alignment CCLs and Mitigations

Deceptive Alignment CCL	Mitigation
Instrumental Reasoning Level 1: the instrumental reasoning abilities of the model enable enough situational awareness (ability to work out and use relevant details of its deployment setting) and stealth (ability to circumvent basic oversight mechanisms) such that, absent additional mitigations, we cannot rule out the model significantly undermining human control.	**Automated monitoring:** a monitoring system to detect illicit use of instrumental reasoning capabilities in a model's chain-of-thought and use it in high-stakes deployments where there is significant risk of the model undermining human control.
Instrumental Reasoning Level 2: the instrumental reasoning abilities of the model enable enough situational awareness and stealth that, even when relevant model outputs (including, e.g., scratchpads) are being monitored, we cannot detect or rule out the risk of a model significantly undermining human control.	***Future work:*** *We are actively researching approaches to addressing models that reach this CCL.*

Safety Cases and Control Evaluations

Looking forward, we expect the approach to deceptive alignment risks to take a similar form as the deployment mitigation process. For each deceptive alignment risk, AI developers should:

- Develop a safety case based on the model's capabilities and the mitigations applied and tested. Such safety cases would make the case that the likelihood of deceptive alignment risk would be low enough for safe deployment, even if the model were trying to meaningfully undermine human control.
- Test the mitigations applied to models via red-teaming protocols called "control evaluations,"[14] to determine whether they meet the requirements of the safety case.

The approach above relies on two safety cases centering respectively on the claims that:

1. A model does not meet a threshold CCL and therefore would be incapable of substantially undermining human control.
2. An automated monitoring process that is validated by a control evaluation is adequate for catching the model's malicious behavior.

Once neither of these apply, that is, once automated monitoring is no longer adequate/can no longer pass a control evaluation, a model is deemed to have reached the second CCL.

Governance and Accountability

Governance

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.

[14] See e.g. https://arxiv.org/abs/2312.06942.

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.

We may also consider disclosing information to other external organizations to promote shared learning and coordinated risk mitigation. We will continue to review and evolve our disclosure process over time.

Future Work

We expect the Framework 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:** While we have updated our CCLs and underlying threat models from version 1.0, there remains significant room for improvement in understanding the risks posed by models in different domains, and refining our set of CCLs.
- **Capability elicitation:** Our evaluators continue to improve their ability to estimate what capabilities may be attainable by different threat actors with access to our models, taking into account a growing number of possible post-training enhancements.
- **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, which may include additional risk domains or higher CCLs within existing domains.
- **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.

Acknowledgements

Version 2.0 of the Frontier Safety Framework was developed by Lewis Ho, Celine Smith, Claudia van der Salm, Joslyn Barnhart, and Rohin Shah, under the leadership of Allan Dafoe, Anca Dragan, Andy Song, Demis Hassabis, Four Flynn, Jennifer Beroshi, Helen King, Nicklas Lundblad, and Tom Lue. The Framework was developed with substantial contributions from Aalok Mehta, Adam Stubblefield, Alex Kaskasoli, Alice Friend, Amy Merrick, Anna Wang, Ben Bariach, Charley Snyder, David Bledin, David Lindner, Dawn Bloxwich, Don Wallace, Eva Lu, Heidi Howard, Iason Gabriel, James Manyika, Joana Iljazi, Kent Walker, Lila Ibrahim, Mary Phuong, Mikel Rodriguez, Peng Ning, Roland S. Zimmerman, Samuel Albanie, Sarah Cogan, Sasha Brown, Seb Farquhar, Sebastien Krier, Shane Legg, Victoria Krakovna, Vijay Bolina, Xerxes Dotiwalla, Ziyue Wang. We are grateful for input from Apollo Research, Carnegie Endowment for International Peace, Center for Long-Term Resilience, Center for the Governance of AI, METR, Redwood Research, UK AI Safety Institute, and several independent researchers.