Frontier Safety Framework

Version 3.0

Published: September 22, 2025

Overview

The Frontier Safety Framework is a set of protocols that aims to address severe risks that may arise from the high-impact capabilities of frontier AI models. It complements Google's suite of AI responsibility and safety practices, and enables AI innovation and deployment consistent with our AI Principles.

The Framework is informed by the broader conversation on Frontier AI Safety and Security Frameworks.[1] The core components of such Frameworks are to:

- Identify capability levels at which frontier AI models, without additional mitigations, could pose severe risk.
- Implement protocols to detect the attainment of such capability levels throughout the model lifecycle.
- Prepare and articulate proactive mitigation plans to ensure severe risks are adequately mitigated when such capability levels are attained.
- Where required or appropriate, involve external parties to help inform and guide the approach.

In the Framework, we specify protocols for the detection of capability levels at which frontier AI models may pose severe risks (which we call "Critical Capability Levels (CCLs)"), and articulate mitigation approaches to address such risks. The Framework addresses misuse risk,[2] risks from machine learning research and development (ML R&D), and misalignment risk.[3] For each type of risk, we define 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 and security of frontier AI models is a global public good. The protocols here represent our current understanding and 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 frontier AI models reaching critical capabilities. These mitigations are most effective when adopted by industry as a whole: our adoption of them would result in effective risk mitigation for society only if all relevant organisations provide similar levels of protection.

The Framework is based on early and evolving research. We may change our approach 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/faisc, https://www.anthropic.com/news/anthropics-responsible-scaling-policy, https://openai.com/index/updating-our-preparedness-framework/, https://www.frontiermodelforum.org/publications/#technical-reports.

[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] Misalignment can pose a number of risks. In the context of the Framework, we address specific scenarios where general-purpose AI agents are potentially misaligned and can become difficult to control, thereby posing a risk of severe harm.

Table of Contents

Section 1: Framework

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

1.1 Scope

The Frontier Safety Framework focuses on possible severe risks stemming from high-impact capabilities of frontier AI models. The Framework complements Google's suite of AI responsibility and safety practices, which address other risks in addition to the severe risks in scope of the Framework in accordance with our AI Principles. The approaches and mitigations outlined in the Framework are not exclusive to models where we believe a severe risk could arise and are part of Google's comprehensive AI responsibility and safety practices.

1.2 Critical Capability Levels

The Framework is built around capability thresholds called "Critical Capability Levels (CCLs)." These are capability levels at which, absent mitigation measures, frontier AI models or systems may pose heightened risk of severe harm. CCLs are determined by identifying and analyzing the main foreseeable paths through which a model could result in severe harm: we then define the CCLs as the minimal set of capabilities a model must possess to do so.

We describe three sets of CCLs: misuse CCLs, machine learning R&D CCLs, and misalignment CCLs.

For **misuse risk**, we define CCLs in the following risk domains where the misuse of model capabilities may result in severe harm:

- **CBRN:** Risks of models assisting in the development, preparation, and/or execution of a chemical, biological, radiological, or nuclear ("CBRN") threat.
- **Cyber:** Risks of models assisting in the development, preparation, and/or execution of a cyber attack.
- **Harmful Manipulation:** Risks of models with high manipulative capabilities potentially being misused in ways that could reasonably result in large scale harm.

For **machine learning R&D risk**, we define CCLs that identify when ML R&D capabilities in our models may, if not properly managed, reduce society's overall ability to manage AI risks. Such capabilities may serve as a substantial cross-cutting risk factor for other pathways to severe harm.

For **misalignment risk**, we outline an exploratory approach that 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.

Most CCLs define one important component of our risk acceptance criteria. Because the CCLs for misalignment risk are exploratory and intended for illustration only, we do not associate them with explicit risk acceptance criteria.

1.3 Outline of Our Risk Assessment Process

For each risk domain, we conduct aspects of our risk assessment at various moments throughout the model development process, both before and after deployment. We conduct a risk assessment for the first external deployment of a new frontier AI model. For subsequent versions of the model, we conduct a further risk assessment if the model has meaningful new capabilities or a material increase in performance, until the model is retired or we deploy a more capable model. The reason for this is because a material change in the model's capabilities may mean that the risk profile of the model has changed or the justification for why the risks stemming from the model are acceptable has been materially undermined.

To identify meaningful new capabilities or material increases in performance, we conduct model capability evaluations, including our automated benchmarks. These evaluations are primarily aimed at understanding the capabilities of the model and may be triggered, for example, upon the completion of a pre-training or post-training run, on various candidates of a model version. These evaluations include a broad range of areas, including general capability evaluations, model behavior, efficiency, coding capabilities, multilinguality, or reasoning. Data from these evaluations are collected and analyzed to give us an indication as to how the model is performing and whether a risk assessment is necessary.

At a high level, our risk assessment involves the following steps (which do not need to be repeated where a previous risk assessment is still appropriate):

- **Identification:** As explained above, we have identified risk domains where, based on early research, we have determined severe risks may be most likely to arise from future models: CBRN, cyber, harmful manipulation, and machine learning R&D.[4] As part of our broader research into frontier AI models, we continue to assess whether there are other risk domains where severe risks may arise and will update our approach as appropriate. For each of the four identified domains, we have developed specific scenarios in which these risks could materialize.
- **Analysis:** Central to our model evaluations are "early warning evaluations," to assess the proximity of the model to a CCL. We define "alert thresholds" for these evaluations that are designed to flag when a CCL may be reached before a risk assessment is conducted 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. We conduct further analysis, including reviewing model independent information, external evaluations, and post-market monitoring as appropriate.
- **Acceptance determination and mitigations:** We then determine whether the model has met or will meet a CCL and, if so, whether we need to implement any further mitigations to reduce the risk to an acceptable level (see below).

Our approach to model evaluations and risk assessments described above means we can proactively monitor a model's capabilities throughout the entire lifecycle of the model and ensure that any severe risk is properly identified and mitigated. Where appropriate, we may engage relevant and appropriate external actors, including governments, to inform our responsible development and deployment practices.

Note on Machine Learning R&D CCLs: Risk assessment must take into account the fact that other actors may put significantly more effort into eliciting capabilities than we put into assessing risk, thus requiring conservatism in the form of evaluations. However, as a frontier AI company, we do not expect other groups to put significantly more effort into ML R&D than we do ourselves. As a result, to assess the ML R&D CCLs, we may use sources of information about our own progress at accelerating ML R&D to assess whether we are near or at the CCLs, in addition to evaluations of ML R&D capabilities. Similarly,

[4] We exclude misalignment risk from this list of domains because of its exploratory nature.

our alert threshold may be defined based on these sources of information, rather than on evaluation scores.

1.4 Response Plans and Mitigations

We apply safety and security mitigations throughout the lifecycle of our models, including as part of our training and model development phase and, where appropriate, before CCLs are reached as described in the process below.

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. 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 are required.

Central to most response plans will be the application of the mitigations described later in this document. We have two categories of mitigations: *security mitigations* (such as preventing the exfiltration of model weights), and *deployment mitigations* (such as safety fine-tuning and monitoring and response) intended to counter the misuse or misaligned expression of critical capabilities in deployments. 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.

1.5 Evaluating Mitigations

We will use various processes to evaluate the effectiveness and limitations of mitigations:

- **Security mitigations**: security infrastructure at Google is subject to penetration testing and other kinds of assessments, and is continually improved based on these results.
- **Deployment mitigations**: we will use a combination of threat modeling, empirical testing, and other sources of information to assess the effectiveness and limitations of our deployment mitigations. These will form the basis of a safety case[5] for models reaching CCLs, that will be reviewed before deployment. See the deployment mitigations sections below for misuse and machine learning R&D for more.

1.6 Summary of Risk Acceptance Criteria

The Framework outlines a variety of risk acceptance criteria for different model risks and capabilities pertaining to severe risks. To summarize:

- A model for which risk assessment indicates no CCL is reached will be deemed to pose an acceptable level of severe risk for further development and deployment, because it should not possess the capabilities required to substantially contribute to severe risk scenarios.[6]
- A model for which the risk assessment indicates a misuse CCL has been reached will be deemed to pose an acceptable level of risk for further development or deployment, if, for example:

[5] A safety case is an assessable argument showing how severe risks associated with a model's CCLs have been reduced to an appropriate level. See also, for reference, https://arxiv.org/abs/2505.01420.

[6] Note that this does not mean that no mitigations will be implemented prior to the deployment of the model. We apply safety and security mitigations throughout the lifecycle of our models for a whole host of potential risks, including as part of our training and model development phase. This section is limited to those mitigations that will be required to bring a model to an acceptable level of risk relative just to the specific CCL, which by definition pose the greatest risks of severe harm.

    - We assess that the deployment mitigations have brought the risk of severe harm to an appropriate level proportionate to the risk, based on considerations such as whether the risk has been reduced to an acceptable level by mitigations, the scope of the deployment, what capabilities and mitigations are available on other publicly available models (e.g. if other models are similarly capable and have few mitigations, then the marginal risk added by our release is likely low), and the historical incidence and severity of related events. This is required only for external deployment, not further development.
    - Security mitigations have been applied to the model weights reaching the recommended security level stated below, or we otherwise assess that the level of security applied is adequate, e.g. if they match or exceed the level of security applied to other models with similar capabilities or risk profiles, or we assess that the benefits of the open release of model weights outweigh the risks.
- A model for which the risk assessment indicates a machine learning R&D CCL has been reached will be deemed to pose an acceptable level of risk for further development or deployment, if, for example:
    - We assess that the deployment mitigations have brought the risk of severe harm to an appropriate level proportionate to the risk, based on considerations such as whether the risk has been reduced to an acceptable level by mitigations, and information pertaining to model propensities and the severity of related events.[7] This is required only for external deployment and large scale internal deployment, not further development.
    - Security mitigations have been applied to the model weights reaching the recommended security level stated below, or we otherwise assess that the level of security applied is adequate, e.g. if they match or exceed the level of security applied to other models with similar capabilities or risk profiles, or we assess that the benefits of the open release of model weights outweigh the risks.
- Because the CCLs for misalignment risk are exploratory and intended for illustration only, we do not associate them with explicit risk acceptance criteria.

Note: Assessing frontier AI capabilities and corresponding severe risk is a complex process. Because the science of AI risk assessment is still developing, our assessments will often involve some level of subjective analysis. The concept of proportionality is central to our determination of whether a particular mitigation has sufficiently reduced the risk to acceptable levels. The mitigation and the effects of such mitigation should also be assessed holistically and be commensurate with expected impact of a model's risk, thus balancing safety with innovation.

[7] Note that deployment mitigations for the ML R&D CCLs are different than those for the misuse CCLs because of differences in which deployment could lead to severe harm for those respective categories. In contrast, security mitigations protecting against weights exfiltration are generally beneficial to both kinds of risk.

Section 2: 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 (CBRN, cyber and harmful manipulation), as well as the mitigation approach that we assess as appropriate for them.

2.1 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 external deployments. For security, we have several levels of mitigations, allowing calibration of the appropriateness and 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. This structured process for deployment mitigations is centered on assessing and reviewing that the risk of severe harm has been brought to an appropriate level proportionate to the risk.

2.1.1 Security Mitigations

Security mitigations against exfiltration risk, such as identity and access management practices and hardening interface-access to unreleased model parameters, 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 use security levels that indicate goals/principles in line with the corresponding level in the RAND framework.[8] Because AI security is an area of active research, we expect the concrete measures implemented to reach each level of security to evolve substantially.

2.1.2 Deployment Mitigations

The following mitigation process for external deployments will be applied to models reaching a misuse CCL, allowing for iterative and flexible tailoring of mitigations to each risk and use case.

1. **Development and assessment of mitigations:** safeguards and an accompanying safety case are developed by iterating on the following:
    a. Developing and improving a suite of safeguards targeting the capability, which may include measures such as safety post-training, monitoring and analysis, account moderation, jailbreak detection and patching, user verification, and bug bounties.[9]
    b. Assessing the robustness of these mitigations against the risk posed through testing (e.g. automated evaluations, red teaming) and threat modeling research. The assessment takes the form of a safety case, and could take into account factors such as:
        i. How much the risk has been reduced by mitigations. For example, whether tests run on mitigated models suggest that the refusal rate and jailbreak robustness together imply the risk has been brought substantially lower than that posed by a model reaching the CCL without mitigations.

[8] In other words, "security level *N*" indicates security controls and detections at a level generally aligned with RAND SL *N*. See https://www.rand.org/pubs/research_reports/RRA2849-1.html, pp 21-22. In aligning our security levels with RAND's, we are referring to the security goals and principles in the RAND framework, rather than the benchmarks (i.e. concrete measures) also described in the RAND report. 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."

[9] See section 5 of https://arxiv.org/abs/2504.01849.

ii. The likelihood and consequences of model misuse, capability improvements after the risk assessment, and likelihood and consequences of our mitigations being circumvented, deactivated, or subverted.
iii. The scope of the deployment. For example, small scale and private deployments may pose substantially less risk than large scale or public deployments.
iv. What capabilities and mitigations are available on other publicly available models. For example, whether another (non-Google) publicly deployed model is at the same CCL, and has mitigations that are less effective at preventing misuse than that of the model being assessed, in which case the deployment of this model is less likely to materially increase risk.
v. The historical incidence and severity of related events: for example, whether data suggests a high (or low) likelihood of attempted misuse of models at the CCL. Mitigations would consequently have to be stronger (or would not have to be so strong) for deployment to be appropriate.

2. **Pre-deployment review of safety case:** external deployments of a model take place only after the appropriate governance function determines the safety case regarding each CCL the model has reached to be adequate. In particular, we will deem deployment mitigations adequate if the evidence suggests that for the CCLs the model has reached, the increase in likelihood of severe harm has been reduced to an acceptable level.
3. **Post-deployment processes:** our safety cases and mitigations may be updated if deemed necessary by post-market monitoring. Material updates to a safety case will be submitted to the appropriate governance function for review.

This process is designed to ensure that residual risk remains at acceptable levels: evidence of efficacy collected during development and testing, as well as expert-driven estimates of other parameters, will enable us to assess residual risk and to detect substantial changes that invalidate our risk assessment. With iteration on safeguards and safety cases, we believe that we are able to make informed decisions about the level of risk via a CCL before a model is released, and reliably prevent models posing unacceptable levels of risk from being deployed.

2.2 Misuse Critical Capability Levels

The table below details a set of CCLs we have identified through ongoing analysis of the CBRN, cyber, and harmful manipulation risk domains. We expect these to evolve over time. We recommend a security level for each of these CCLs, which reflect our assessment of the minimum appropriate level of security 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 proportionate to the risks posed and may be adjusted if our understanding of the risks changes. This may occur if, for example, a model does not possess capabilities meaningfully different from other publicly available models that have weaker security applied (in which case the marginal benefit of higher security is limited), 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.

2.2.1 Chemical, Biological, Radiological or Nuclear

This risk domain focuses on risks of models assisting in the development, preparation and/or execution of a CBRN threat.

Table 2.2.1.a: CBRN CCLs and Security Mitigations

Critical Capability Level	Recommended security level and rationale
CBRN uplift level 1: Provides low to medium resourced actors uplift in reference scenarios resulting in additional[10] expected harm at severe scale.	**Security level 2**[11] The difficulty of building defenses against certain CBRN threats means the exfiltration and leak of model weights with this capability could be highly damaging. However, the low to medium resourced actors who would be likely to experience the most CBRN uplift are unlikely to pose a substantial exfiltration threat at the level of RAND OC3 groups.

2.2.2 Cyber

This risk domain focuses on risks of models assisting in the development, preparation, and/or execution of a cyber attack.

Table 2.2.2.a: Cyber CCLs and Security Mitigations

Critical Capability Level	Recommended security level and rationale
Cyber uplift level 1: Provides sufficient uplift with high impact cyber attacks for additional expected harm at severe scale.	**Security level 2** Models able to greatly assist cyber attack might be of interest to well-resourced state actors. However, the potential for automated cyber-defense and social adaptation as a response to exfiltration means that higher levels of security, and the resulting costs to innovation, are likely not warranted.

2.2.3 Harmful Manipulation

This risk domain focuses on risks of models with high manipulative capabilities potentially being misused in ways that could reasonably result in large scale harm.

Note: The research into harmful manipulation from a severe risk perspective is nascent. The CCL and our assessment of the risk in this domain is exploratory and subject to further research, and may be substantially changed over time.

Table 2.2.3.a: Harmful Manipulation CCLs and Security Mitigations

Critical Capability Level	Recommended security level and rationale
Harmful manipulation level	**Security level 2**

[10] Here, and in other misuse CCLs, we intend this to mean relative to a baseline without generative AI.

[11] Mitigations at this level may include model access management, physical security controls, authentication measures, endpoint security, access management, secure model storage, vulnerability detection & management, detection of & response to suspected malicious activity.

1 (exploratory): Possesses manipulative capabilities sufficient to enable it to systematically and substantially change beliefs and behavior in identified high stakes contexts over the course of interactions with the model, reasonably resulting in additional expected harm at severe scale.	The lower velocity of harm scenarios associated with this CCL and the viability of social defenses against large scale misuse of such models count against security mitigations with substantial costs to innovation.

Section 3: Machine Learning R&D

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

3.1 Mitigation Approach

As with misuse CCLs, we designate security and deployment mitigations for ML R&D CCLs, although deployment mitigations focus on different threat models and therefore also include measures for large scale internal deployments.

3.1.1 Security Mitigations

Security mitigations against exfiltration risk are important for models reaching ML R&D CCLs, because the exfiltration of highly capable models increases the likelihood they will be misused to achieve other critical capabilities, or deployed without adequate control and oversight. Below, we rely again on security levels to articulate the security goal recommended for each CCL.[12] Security mitigations also protect against the risk of the model exfiltrating itself.

3.1.2 Deployment Mitigations

The following mitigation process will be applied for deployments of a model reaching a ML R&D CCL. The approach is similar to misuse deployment mitigations, with an added focus on large scale internal deployments.

1. **Development and assessment of mitigations:** safeguards and an accompanying safety case are developed by iterating on the following:
    a. Developing and improving a suite of safeguards targeting the capability, which may include measures such as limiting affordances, monitoring and escalation, auditing, and alignment training, in addition to measures for preventing large scale misuse.[13]
    b. Assessing the robustness of these mitigations against the risk posed in both internal and external deployment through testing (e.g. automated evaluations, red teaming) and threat modeling research. The assessment takes the form of a safety case, taking into account factors such as:
        i. How much the risk has been reduced by mitigations. For example, tests run on the safeguards may suggest that it is very unlikely they can be circumvented by external threat actors or the model in question to increase ML R&D risk.
        ii. The likelihood and consequences of model misuse or misalignment, capability improvements after the risk assessment, and likelihood and consequences of our mitigations being circumvented, deactivated, or subverted.
        iii. The scope of the deployment. For example, small scale and private deployments may pose substantially less risk than large scale or public deployments.
        iv. Model propensity for, historical incidence of and severity of related events: for example, such data may suggest a high (or low) likelihood of misalignment in or misuse of models at the CCL, and mitigations would consequently have to be stronger (or not as strong) for deployment to be appropriate.
2. **Pre-deployment review of safety case:** external deployments and large scale internal deployments of a model take place only after the appropriate governance function determines

[12] While the RAND framework is not specifically designed to address this case, we index it at present because it is the most useful reference in this area.

[13] See section 6 of https://arxiv.org/abs/2504.01849.

the safety case regarding each CCL the model has reached to be adequate. In particular, we will deem deployment mitigations adequate if the evidence suggests that for the CCLs the model has reached, the increase in likelihood of severe harm has been reduced to an acceptable level.

3. **Post-deployment processes:** our safety cases and mitigations may be updated if deemed necessary by post-market monitoring. Material updates to a safety case will be submitted to the appropriate governance function for review.

This process is designed to ensure that residual risk remains at acceptable levels: evidence of efficacy collected during development and testing, as well as expert-driven estimates of other parameters, will enable us to assess residual risk and to detect substantial changes that invalidate our risk assessment. With iteration on safeguards and safety cases, we believe that we are able to make informed decisions about the level of risk via a CCL before a model is released or deployed internally at scale, and reliably prevent models posing unacceptable levels of risk from being deployed.

3.2 ML R&D Critical Capability Levels

The table below details a set of ML R&D CCLs we have identified that may lead to heightened severe risk through ML R&D. We expect these to evolve over time. We recommend a security level for each of these CCLs, which reflect our assessment of the minimum appropriate level of security the field of frontier AI should apply to models reaching each CCL.[14]

3.2.1 Machine Learning R&D

These CCLs focus on risks posed by models capable of accelerating the rate of AI progress. These capabilities may indicate a heightened ability to undermine human control of models, may incentivize greater (and therefore higher risk) deployment of models, and could also result in the unsafe attainment or proliferation of other powerful AI models if misused by external threat actors.

Table 3.2.1.a: Machine Learning R&D CCLs and Security Mitigations

Critical Capability Level	Recommended security level and rationale
ML R&D acceleration level 1: Has been used to accelerate AI development, resulting in AI progress substantially accelerating from historical rates.	**Security level 3**[15] Unrestricted access to models at this level of capability could significantly increase a threat 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 models, effects that may have long-lasting consequences. Substantially strengthened security is therefore recommended. However, we expect that acceleration will stem from systems of models integrated with workflows, rather than the model alone. The overall reduced impact of model weights counts against security levels with substantial innovation costs.

[14] The same caveats regarding security levels for misuse CCLs apply.

[15] This level may include mitigations aligned with SL 2, plus additional mitigations designed to prevent unilateral access, harden infrastructure, and prevent data exfiltration.

ML R&D automation level 1: Can fully automate the work of any team of researchers at Google focused on improving AI capabilities, with approximately comparable all-inclusive costs.	**We recommend Security level 4[16] 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 (or AI systems) 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.

[16] This level may include mitigations aligned with SL 2 and 3, plus additional mitigations aimed to isolate model weights, enhanced data center security, further hardening of infrastructure and minimizing potential attack surface.

Section 4: Misalignment (Exploratory Approach)

Misalignment can pose a number of risks. In the context of the Framework, we address specific scenarios where general-purpose AI agents are potentially misaligned and can become difficult to control, thereby posing a risk of severe harm. Here we describe an approach for addressing misalignment risk that focuses specifically on when models may develop a baseline instrumental reasoning ability at which, without additional mitigations, they may have the potential to undermine human control. When models reach this capability level, one possible mitigation is to apply an automated monitor to the model's explicit reasoning (e.g. chain-of-thought output). 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.

Given its nascency, we expect our approach to misalignment risk to evolve substantially. This section is therefore illustrative only. Accordingly, we do not indicate security mitigations for models at these CCLs. The table below summarizes the initial approach:

Table 4.a: Exploratory Misalignment CCLs and Mitigations

Illustrative Instrumental Reasoning CCL	Illustrative 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:** monitoring system to detect illicit use of instrumental reasoning capabilities in a model's chain-of-thought may be applied in high-stakes internal 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.*

Section 5: Updates and Disclosures

5.1 Updates

The Frontier Safety Framework will be updated at least once a year—more frequently if we have reasonable grounds to believe the adequacy of the Framework or our adherence to it has been materially undermined. The process will involve (i) an assessment of the Framework's appropriateness for the management of systemic risk, drawing on information sources such as record of adherence to the framework, relevant high-quality research, information shared through industry forums, and evaluation results, as necessary, and (ii) an assessment of our adherence to the Framework. Following this assessment, we may:

- Update our risk domains and CCLs, where necessary.
- Update our testing and mitigation approaches, where needed to ensure risk remains adequately assessed and addressed according to our current understanding.

The updated version and framework assessment will be reviewed by the appropriate corporate governance bodies.

5.2 Disclosures

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 relevant information with appropriate government authorities where it will facilitate safety of frontier 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 organisations to promote shared learning and coordinated risk mitigation. We will continue to review and evolve our disclosure process over time.

5.3 Past Updates and Changes

Past versions:

- Version 2.0 (4 February 2025)
- Version 1.0 (17 May 2024)