Frontier Safety Framework

Version 3.1

Published: April 17, 2026

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 Google's 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 significant or severe risks (which we call "Tracked Capability Levels (TCLs)" and "Critical Capability Levels (CCLs)" respectively), and articulate mitigation approaches to address such risks. The Framework addresses misuse risk[2] as well as machine learning research and development (ML R&D) and misalignment risk.[3] For each type of risk, we define a set of CCLs (and TCLs where relevant) and a mitigation approach for them. Risk assessment will necessarily involve evaluating cross-cutting capabilities such as agency, tool use, reasoning, and scientific understanding. Please refer to the Glossary at the end of this document for definitions used in the Framework.

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 organizations 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/rsp-updates, https://www.anthropic.com/news/compliance-framework-SB53, 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] In the context of the Framework, we address specific scenarios where model misalignment may create the risk of significant or severe harm, and scenarios where ML R&D capabilities may heighten misalignment, misuse, and structural risks.

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 and Tracked Capability Levels

The Framework is built primarily 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. CCLs are one important component of our risk acceptance determination.

We identify CCLs for two kinds of risks: misuse risk and risks related to machine learning R&D and misalignment.

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 and misalignment risks**, 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 several pathways (e.g. through misalignment, misuse or structural risks) to severe harm.

This update to the Framework introduces **"Tracked Capability Levels (TCLs)."** TCLs are meant to capture significant risks that may manifest at a lower capability threshold than our CCLs and we apply our mitigation and risk acceptance determination process proportionately. We identify TCLs for CBRN risk, as well as ML R&D and misalignment risks. We may include TCLs for additional risks in the future, as our threat modeling develops. Similarly to CCLs, early warning evaluations will be used to assess the proximity of a model to a TCL and analyze the risk posed, involving internal and external experts as needed.

1.3 Risk Management Process

We conduct our risk management process as appropriate, throughout the model development process, on various checkpoints or versions of a model, both before and after deployment.

1.3.1 Risk Identification

The first part of our risk management process is **risk identification**. We identify potential risks that could stem from our models and analyze their characteristics to determine which of the identified risks could be significant or severe risks. We consider a wide range of risks as part of our ongoing research, taking into account the characteristics, capabilities, propensities, and affordances of our models and other sources of information, such as our internal risk taxonomies, internal expertise and relevant external research. As explained above, we have identified risk domains where, based on early research, we have determined significant or severe risks may be most likely to arise from future models: CBRN, cyber, harmful manipulation, as well as machine learning R&D and misalignment. As part of our broader research into and development of frontier AI models, we continue to assess whether there are other risk domains where significant or severe risks may arise and will update our approach as appropriate. For each of the four identified domains, we have developed specific scenarios and T/CCLs in which these risks could materialize.

1.3.2 Inherent Risk Assessment

To understand whether a model may, without appropriate mitigations, contribute to the risks identified, we conduct a closer assessment of its capabilities against the T/CCLs. We conduct a **critical capability assessment** prior to the first external deployment[4] of a new frontier AI model. For external deployment of subsequent versions of the model, we determine whether a substantial modification has been made and whether a further critical capability assessment is required if (1) the model has meaningful new capabilities or material increases in performance; and (2) we believe such material capability increase could materially undermine the justification for why the risks stemming from the model are acceptable.

To understand if a subsequent version of the model has meaningful new capabilities or material increases in performance relative to the last checkpoint subject to a critical capability assessment, we conduct **material capability change assessments** on various checkpoints upon the completion of a post-training run. These assessments draw on model characteristics, such as its provenance, size, modality, and performance on general capability benchmarks. To understand if such a change in capability in the subsequent version of the model could materially undermine the justification for why the risks stemming from the model are acceptable, we may also conduct light-weight versions of our "early warning evaluations" described below, including our automated benchmarks, to confirm a model remains below T/CCL. We may also apply material capability change assessments to models not on the frontier, to understand whether a critical capability assessment is required.

Central to our critical capability assessments are "early warning evaluations," which we use to test the specific threats and risk scenarios identified through our threat modeling, determine a model's capability, and assess the proximity of the model to a T/CCL. For CCLs, we define "alert thresholds," which may draw on evaluation results, expert assessments, and other sources of information; that are designed to flag when a CCL may be reached before a critical capability assessment is conducted again. In our evaluations, we seek to apply appropriate scaffolding, inference compute, and other augmentations to also assess the capabilities of systems that will likely be produced with the model based on the specific threats and risk scenarios we have identified for the T/CCL. We may run early warning evaluations more

[4] A critical capability assessment may not be conducted for low-risk external deployments (e.g. to a small number of trusted testers) if the appropriate governance function determines the residual risk of such deployments to be acceptable even if the model has reached a T/CCL (including with a safety case where a CCL has been reached).

frequently or adjust alert thresholds 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. In particular, we strive to learn from our post-market monitoring mechanisms, including as part of our detection, mitigation, response and/or reporting of incidents relating to our frontier safety risk domains. Actionable insights from these processes allow us to enhance our tools, training, processes, policies, and response efforts.

Our approach to model evaluations and inherent risk assessments described above means we can proactively monitor a model's capabilities throughout the entire lifecycle of the model and ensure that any significant or severe risk is properly identified and mitigated. Where appropriate, we may engage relevant 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 ML R&D CCLs, in addition to evaluations of ML R&D capabilities.

1.3.3 Risk Mitigation

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 T/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. Central to most response plans will be the application of the mitigations described later in the Framework. However, if we assess that risks stemming from the model remain acceptable and model capabilities remain distant from a CCL, the response plan may include updating the alert threshold to ensure the safety buffer remains appropriate.

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.

The mitigations described in this document are limited to those mitigations that will be required to bring a model to an acceptable level of risk relative just to the specific T/CCL, which by definition pose the greatest risks of harm. Because we cannot always anticipate what security and deployment mitigations will be appropriate for models beyond the current frontier, the specific mitigations we implement may be determined when a T/CCL is reached, informed by the threat landscape at that time. These mitigations also reflect considerations from the perspective of addressing significant and 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.

1.3.4 Residual Risk Assessment

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.

Models reaching TCLs or CCLs will be subject to residual risk assessments as part of evaluating their deployment mitigations. For models reaching CCLs, the residual risk assessment will be informed by a supplemental safety case. See the deployment mitigations sections below regarding misuse and machine learning R&D and misalignment for more.

1.3.5 Risk Acceptance Determination

The Framework outlines the risk acceptance determination approach for different model risks and capabilities pertaining to significant and severe risks. To summarize:

- A model for which inherent risk assessment indicates no T/CCL is reached will be deemed to pose an acceptable level of risk for further development and deployment, because it should not possess the capabilities required to substantially contribute to significant or severe risk scenarios.
- A model for which the inherent risk assessment indicates a misuse T/CCL has been reached will be deemed to pose an acceptable level of residual risk for further development or deployment, if:
    - We assess that the deployment mitigations have brought the residual risk of harm to an acceptable level, based on considerations such as the effectiveness of 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 external deployment is likely low), and the historical incidence and severity of related events. This is required only for external deployment, not internal deployment or 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. based on mitigations already in place, 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 inherent risk assessment indicates a ML R&D CCL has been reached will be deemed to pose an acceptable level of residual risk for further development or deployment, if:
    - We assess that the deployment mitigations have brought the residual risk of severe harm to an acceptable level, based on considerations such as the effectiveness of mitigations, and information pertaining to model propensities and the severity of related events.[5] This is required only for external deployment and high-risk 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. based on mitigations already in place, 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.

[5] 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 and unauthorized modification are generally beneficial to both kinds of risk.

- A model for which the inherent risk assessment indicates the TCL for ML R&D and misalignment risk has been reached will be deemed to pose an acceptable level of residual risk for further development or deployment, if:
    - We assess that the deployment mitigations have brought the residual risk of significant harm to an acceptable level, based on considerations such as the effectiveness of mitigations, and information pertaining to model propensities and the severity of related events, or if we assess that the model is very unlikely to possess the propensities that would be required for material risk of significant harm through misalignment risk. This is required only for external deployment and high-risk internal deployment, not further development.
    - We assess that the level of security applied is adequate, e.g. based on mitigations already in place, 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.

Note: Assessing frontier AI capabilities and corresponding significant and 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 proportionate with expected impact of a model's risk, thus balancing safety with innovation.

Section 2: Misuse

This section describes our mitigation approach for models that pose risks of significant or severe harm through misuse, and then details our set of misuse T/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 or unauthorized modification 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. Regarding deployment mitigations for models reaching T/CCLs, we specify a standard process for applying, assessing, and reviewing mitigations: the aim of this process is to calibrate mitigations to T/CCLs, and the procedural approach reflects the more iterative and T/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 acceptable level proportionate to the risk.

2.1.1 Security Mitigations

Security mitigations against exfiltration or unauthorized modification risk, such as identity and access management practices and hardening interface-access to unreleased model parameters, are important for models reaching misuse CCLs. This is because the release or unauthorized modification 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.

Using Google's Secure AI Framework (SAIF) and Google's common security infrastructure, we implement state-of-the-art security mitigations. SAIF is a defense-in-depth approach that embeds security into every layer of our AI systems—from data and infrastructure to models and applications. SAIF is premised on six core elements: strong security foundations, detection and response, automated defenses, harmonized platform-level controls, adaptable controls to mitigate emerging threats, and end-to-end risk assessments.

We use security levels to indicate security goals/principles in line with the corresponding level in the RAND model weight security framework.[6] We also define and recommend "Security Level 2+," which uses RAND Security Level 2 (SL2) as a baseline, with additional security measures designed to address risks from insider threats and well-resourced non-state external actors. These additional measures may include, for example: dedicated insider risk teams; background checks and ID verification for personnel with sensitive access; review of model training data for signs of tampering; mandating that the processing of untrusted inputs occurs within sandboxed environments; advanced red-teaming that simulates well-resourced adversaries (including Advanced Persistent Threat (APT) groups); and proactive threat hunting with 24/7 incident response capabilities.

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

[6] 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."

2.1.2 Deployment Mitigations

The following mitigation process for external deployments will be applied to models reaching a misuse T/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 residual risk assessment 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.[7]
    b. Assessing the robustness of these mitigations against the risk posed through testing (e.g. automated evaluations, red teaming) and threat modeling research. This residual risk assessment could take into account factors such as:
        i. The effectiveness of the mitigations. For example, tests run on mitigated models may suggest that the refusal rate and jailbreak robustness together imply that threat actors are unlikely to be able to circumvent safeguards.
        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 T/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 T/CCL. Mitigations would consequently have to be stronger (or would not have to be so strong) for deployment to be appropriate.
    c. Where the model has reached a CCL, the residual risk assessment will be supplemented with a safety case.
2. **Pre-deployment review of residual risk assessment:** external deployments of a model take place only after the appropriate governance function determines the residual risk to be acceptable (including a safety case where a CCL has been reached). In particular, we will deem deployment mitigations adequate if the evidence suggests that for the T/CCLs the model has reached, the increase in likelihood of harm from the proposed external deployment has been reduced to an acceptable level.
3. **Post-deployment processes:** our residual risk assessments, safety cases and mitigations may be updated as a result of post-market monitoring, including information about incidents relating to our frontier safety risk domains. Material updates to a safety case will be submitted to the appropriate governance function for review and might result in updates to the related residual risk assessment or safety case.

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 mitigations and residual risk assessments, we believe that we are able to make informed decisions about the level of risk via a T/CCL before a model is deployed, and prevent models posing unacceptable levels of risk from being deployed.

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

While we monitor for potential future risks related to insider misuse, our current assessment of these risk domains indicates that the risks of insider misuse posed by the misuse T/CCLs identified below are effectively addressed by our baseline controls (such as background checks and access restrictions). At this stage, additional mitigations beyond these established safeguards are not required.

2.2 Misuse Capability Levels

We set out below a set of CBRN, cyber, and harmful manipulation CCLs as well as the CBRN TCL that we have identified through ongoing analysis of these risk domains. We expect these to evolve over time. We recommend a security level for each CCL, which reflects 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 CCL 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[8] expected harm at severe scale.	**Security level 2+** The difficulty of building defenses against certain CBRN threats means the exfiltration, unauthorized modification, 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 OC3 groups as defined by RAND.

CBRN Tracked Capability Level

The CBRN TCL is defined as follows:

> **CBRN uplift TCL**: Provides low to medium resourced actors uplift in reference scenarios, resulting in material risk of CBRN attacks that could result in significant but not severe harm.

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

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 CCL 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 CCL and Security Mitigations

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

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

3.1 Mitigation Approach

As with misuse T/CCLs, we take a similar approach to security and deployment mitigations for ML R&D and Misalignment T/CCLs, although deployment mitigations focus on different threat models and therefore also include measures for high-risk internal deployments.

3.1.1 Security Mitigations

Security mitigations against exfiltration and unauthorized modification risk are important for models reaching ML R&D CCLs. In addition, 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.[9] 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 and Misalignment T/CCL. The approach is similar to misuse deployment mitigations, with an added focus on high-risk internal deployments.

1. **Development and assessment of mitigations:** safeguards and an accompanying residual risk assessment 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.[10]
    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. This residual risk assessment could take into account factors such as:
        i. The effectiveness of the 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.
        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 T/CCL, and mitigations would consequently have to be stronger (or not as strong) for deployment to be appropriate.
    c. Where the model has reached a CCL, the residual risk assessment will be supplemented with a safety case.

[9] 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.
[10] See section 6 of https://arxiv.org/abs/2504.01849.

2. **Pre-deployment review of residual risk assessment:** external deployments and high-risk internal deployments of a model take place only after the appropriate governance function determines the residual risk to be acceptable (including a safety case where a CCL has been reached). In particular, we will deem deployment mitigations adequate if the evidence suggests that for the T/CCLs the model has reached, the increase in likelihood of harm from the proposed deployment has been reduced to an acceptable level.
3. **Post-deployment processes:** our residual risk assessments, safety cases and mitigations may be updated as a result of post-market monitoring, including information about incidents relating to our frontier safety risk domains. Material updates to a safety case will be submitted to the appropriate governance function for review and might result in updates to the related residual risk assessment or safety case.

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 mitigations and residual risk assessments, we believe that we are able to make informed decisions about the level of risk via a T/CCL before a model is deployed, and prevent models posing unacceptable levels of risk from being deployed.

3.2 ML R&D and Misalignment Capability Levels

We take a tiered approach to addressing ML R&D and misalignment risks:

- We set a Stealth and Situational Awareness TCL that indicates a baseline level of model capability above which misalignment risks may begin to arise. At this capability level we apply a proportionate set of mitigation and risk acceptance determination processes.
- We set ML R&D CCLs that indicate higher capability levels at which misalignment, misuse and structural risks are heightened and may reach a severe scale. At these capability levels we apply a stronger set of mitigations, and follow the risk acceptance process as detailed in the rest of the framework.

3.2.1 Stealth and Situational Awareness Tracked Capability Level

The Stealth and Situational Awareness TCL is defined as follows:

> **Stealth and Situational Awareness TCL**: The instrumental reasoning abilities of the model enable enough situational awareness (ability to discover 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.

When a model has reached this TCL, we will carry out periodic residual risk assessments of the misalignment risk posed. This assessment may take into account models' alignment propensities, capabilities, and our defenses against misaligned models. Since misalignment risks can manifest through internal use, these assessments will also account for the safety implications of high-risk internal deployments for such models. If the risk assessment deems the residual risk from internal deployment to be unacceptable without additional safeguards (such as chain-of-thought monitoring on high-risk deployments), these will be applied.

3.2.2 ML R&D Critical Capability Levels

We define ML R&D CCLs at capability levels at which misalignment, misuse and structural risks may reach a severe scale. 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.[11]

Table 3.2.2.a: ML 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**[12] 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.
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**[13] **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.

[11] The same caveats regarding security levels for misuse CCLs apply.
[12] This level may include mitigations aligned with SL 2+, plus additional mitigations designed to prevent unilateral access, harden infrastructure, and prevent data exfiltration.
[13] 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: Governance and Accountability

4.1 Governance structure

We have in place a well-established and comprehensive internal governance structure designed to ensure the robust implementation of the processes outlined in this Frontier Safety Framework. Responsibilities for assessing and mitigating risks are clearly defined and allocated across all levels of the organization. This includes legal, compliance, and safety reviews with escalation procedures to ensure appropriate oversight.

Section 5: Updates and Disclosures

5.1 Updates

The Frontier Safety Framework will be reviewed 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 significant and severe 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 T/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 organizations 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

Versions:

- Version 3.1 (April 17, 2026)
    - Introduced CBRN TCLs and outlined mitigation and risk acceptance process.
    - Incorporated the previous exploratory Misalignment risk domain into a combined ML R&D and Misalignment risk domain and outlined mitigation and risk acceptance process for the TCL.
    - Outlined enhanced level of security for CBRN, Cyber and Harmful Manipulation CCLs to Security Level 2+ to protect against non-state actors and insider threats.
    - Included more detail on our risk management process.
    - Included description of our internal governance structure.
    - Introduced a glossary.
- Version 3.0 (September 22, 2025)
- Version 2.0 (February 4, 2025)
- Version 1.0 (May 17, 2024)

Glossary

Model

Frontier AI Models or Models: are trained on a large data set, display significant generality, are capable of performing a wide range of distinctive tasks and have high-impact capabilities. Frontier AI models' agentic and reasoning-based general capabilities near or exceed those of other Google models. There may be many different checkpoints and versions of the same model along the entire model lifecycle: we consider checkpoints and versions the same model if their base capabilities stem primarily from the same foundational training run, but will conduct additional testing on new versions as specified in Section 1.3. When we refer to "model" in this Framework, we may be referring to a checkpoint or version of the model.

Risk Domains

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.

ML R&D and Misalignment: risks of models' ML R&D capabilities or misaligned propensities reducing society's overall ability to manage AI risks. Such capabilities may serve as a substantial cross-cutting risk factor for several pathways (e.g. through misalignment, misuse or structural risks) to severe harm.

Thresholds

Critical Capability Levels (CCLs): are the main capability thresholds around which we have built the Framework process. They represent the capability levels at which, absent mitigation measures, frontier AI models or systems may pose heightened risk of severe harm.

Tracked Capability Levels (TCLs): are capability thresholds which capture a lower level of risks than our CCLs. They represent the capability levels at which, absent mitigation measures, frontier AI models or systems may pose heightened risk of significant but not severe levels of harm.

Alert Thresholds: are thresholds which we set marginally earlier than our CCLs. Crossing these thresholds indicates a CCL may be reached in the foreseeable future, due to the speed of capability progress. In particular, they are meant to account for the possibility of models reaching a CCL before the next critical capability assessment cycle.

Inherent Risk Assessment

Inherent Risk Assessments: the process of evaluating the level of risk posed by the model. This is part of the risk management process, and includes material capability change assessments and critical capability assessments.

Critical Capability Assessments: determine a model's proximity to T/CCLs through the use of early warning evaluations and other sources of evidence.

- **Early Warning Evaluations:** are evaluations which measure the dangerous capabilities of a model. They specifically target the threats and risk scenarios identified through our threat

modeling to determine a model's proximity to a CCL or TCL, and they provide the most direct evidence for whether these thresholds have been reached.

Material Capability Change Assessments: indicate whether a critical capability assessment is required. These assessments are typically conducted upon the completion of new post-training runs on various checkpoints, and they are primarily aimed at determining whether updated versions of a model have developed *material capability increases* relative to the last checkpoint subject to a critical capability assessment. May also be used to assess whether a critical capability assessment is required of non-frontier AI models.

- **Material Capability Increases:** are meaningful new capabilities or material increases in model performance that we believe could materially undermine our justifications for a model's level of risk being acceptable, for example, by reaching a CCL that the previous version of the model had not reached.

Risk Mitigation

Response Plans: are plans put in place when a new alert threshold has been reached to prepare for future models reaching a CCL. Central to most response plans will be the application of the mitigations described earlier in the Framework. However, if we assess that risks stemming from the model remain acceptable and model capabilities remain distant from a CCL, the response plan may include updating the alert threshold to ensure the safety buffer remains appropriate.

Deployment Mitigations: are safety measures we implement which are intended to counter the misuse or misaligned expression of critical capabilities in deployments. They may include safety post-training, input/output/chain-of-thought monitoring and analysis, account moderation, jailbreak detection and patching, user verification, and bug bounties.

- **External Deployments:** represent model releases to entities outside of Google. They can include both *low-risk external deployments* as well as general purpose releases to the broader public.
- **Low-Risk External Deployments:** represent small scale and private releases restricted to particular external groups or organizations.
- **Internal Deployments:** represent model releases restricted to Google employees for internal use.
- **High-Risk Internal Deployments:** represent model releases restricted to Google employees for use cases with the potential to enable severe threat scenarios (e.g. building internal security infrastructure or automating ML R&D).

Security Mitigations: are safety measures we implement which are intended to prevent the unauthorized modification or exfiltration of model weights by unauthorized actors. They may include identity and access management practices, and hardening interface-access to unreleased model parameters, among other measures. We communicate the level of security we consider appropriate to models which cross our thresholds using the RAND model weight security framework.

- **RAND Security Levels:** indicate security goals and principles relevant to frontier developers aiming to protect their model weights from various actors - with the strength and type of security measures differing for each category of actor.[14]

[14] See https://www.rand.org/pubs/research_reports/RRA2849-1.html.

Residual Risk Assessment

Residual Risk Assessments: the process of evaluating the level of risk that remains after all planned mitigation strategies have been implemented. This is part of the risk management process, following a critical capability assessment and the effect of mitigations. The residual risk should be brought within the acceptable levels, as determined by the T/CCLs as relevant.

Safety Cases: are an assessable argument showing how severe risks associated with a model's CCLs have been reduced to an acceptable level. Residual risk assessments will be supplemented with a safety case when a model reaches a CCL, and inform a model's risk acceptance determination.[15] For misuse risk, such an argument may be based, for example, on considerations such as the effectiveness of mitigations, the scope of the deployment, what capabilities and mitigations are available on other publicly available models, and the historical incidence and severity of related events. For ML R&D and misalignment risk, such an argument may be based, for example, on considerations such as the effectiveness of mitigations relative to model capabilities, and information pertaining to model propensities and the severity of related events.

[15] See also, for reference, https://arxiv.org/abs/2505.01420.