Example Safety and Security Framework (Draft)

This document is an example of a Safety and Security Framework, written from the perspective of a hypothetical frontier AI developer aiming to address potential catastrophic risks that might arise from advanced model development and deployment. It draws primarily from the Safety and Security chapter of the EU GPAI Code of Practice and secondarily from "Common Elements of Frontier AI Safety Policies," though may not necessarily fully adhere to the Code of Practice. This document includes both example language and placeholder sections and does not necessarily represent METR's view of an ideal safety framework. We release this into the public domain under the Creative Commons Zero (CC0) license. This document was last updated on August 13, 2025.

Formats: PDF, DOCX. An older version based on the Third Draft Code of Practice can be found here.

Example Safety and Security Framework (Draft)

This document is an example of a Safety and Security Framework, written from the perspective of a hypothetical frontier AI developer aiming to address potential catastrophic risks that might arise from developing, making available, and using advanced models. It draws primarily from the Safety and Security chapter of the EU GPAI Code of Practice and secondarily from "Common Elements of Frontier AI Safety Policies" and California's SB-53 Transparency in Frontier Artificial Intelligence Act, though it may not be fully adherent and compliant. This document includes both example language and placeholder sections. We do not believe it is state-of-the-art, reduces risks to acceptable levels, or represents our recommendations absent regulatory requirements. We release this into the public domain under the Creative Commons Zero (CC0) license. This document was last updated on November 18, 2025.

Formats: PDF, DOCX. The previous version can be found here.

Summary

This Framework outlines our commitments for evaluating and mitigating systemic risks that may arise from frontier AI models that we develop, make available on the market, or use. Systemic risks refer to risks that could reasonably foreseeably cause large-scale harm to public health, safety, public security, fundamental rights, or society at large, as described in Risk identification. The Framework provides a methodology for determining when risks reach unacceptable levels, implementing appropriate technical safety and security mitigations, and ensuring accountability throughout the AI model lifecycle.

Systemic risk assessment

The Framework uses a structured process to identify systemic risks relevant to each model. This involves assessing the four primary risk categories detailed below—cyber offence, CBRN, loss of control, and harmful manipulation—as well as identifying any additional systemic risks specific to the model's capabilities. Each selected risk, and overall systemic risk, is then analyzed to determine whether it is acceptable. If it is not acceptable, we implement appropriate mitigations and then reassess the risk.

For each systemic risk, the Framework:

- Defines systemic risk tiers based on model capabilities,
- Describes the mitigations required at each risk tier,
- Shares how we determine whether the risk is acceptable, and
- Estimates when we may reach higher risk tiers.

Evaluation and measurement methodology

The Framework implements an approach to model evaluation that:

- Defines specific trigger points throughout development and deployment for conducting lighter-touch and full evaluations,
- Requires state-of-the-art elicitation techniques to avoid underestimating model capabilities,
- Maintains scientific standards for internal validity, external validity, and reproducibility,
- Integrates open-ended testing to discover unexpected behaviors and properties, and
- Includes appropriate safety margins.

Key

Derived from Code of Practice (CoP) Measure 1.1(1)–(4) or the EU AI Act

Derived from CoP Measure 1.1¶2:

> *The Framework will contain a high-level description of implemented and planned processes and measures for systemic risk assessment and mitigation to adhere to this Chapter.*

We interpret this requirement to include a description of how the signatory will comply with CoP commitments 2–6, since “system risk assessment” is defined in the glossary to include commitments 2–4, and systemic risk mitigations are in commitments 5–6.

Derived from SB53, Transparency in Frontier Artificial Intelligence Act

Not directly derived from the above and include:

- Example methods for CoP Measures 2–6
- Processes relevant to CoP Measures 7–10
- Best practices, which CoP Commitment 1 requires developers to meet and exceed: “Signatories commit to adopting a state-of-the-art Safety and Security Framework,” where state-of-the-art is “the forefront of relevant research, governance, and technology that goes beyond best practice.”

Decision-making process

The Framework establishes a structured process for determining whether to proceed with development, release, or use:

- Risk assessment results are compared against systemic risk tiers.
- Mitigation effectiveness is evaluated against the measured risk level.
- Safety margins are applied based on uncertainty levels.
- Independent external assessment may inform decisions.
- Development, make available on the market, or use proceeds only when corresponding risk is deemed acceptable.

Technical mitigations

When proceeding with development or deployment, the Framework requires implementing:

- Safety mitigations such as training data filtering, staged model access, and chain-of-thought transparency. Post-deployment monitoring is used to detect unexpected misuse or hazardous capabilities.
- Security mitigations that must be robust, as described below, and must additionally be designed to protect against threat actors. These include protection measures for unreleased model weights including model weight encryption, limiting unmonitored internal access to model weights, and insider threat mitigations. This requirement only applies to models whose capabilities are not inferior to those of at least one open-weight model available on the market.

Accountability and transparency measures

The Framework establishes structures including:

- Clear allocation of systemic risk responsibilities across the organization
- Independent external assessments of models
- Serious incident reporting with defined timelines based on incident severity
- Regular framework assessments (at least annually or upon material changes)
- Public transparency through publication of framework and model report summaries as necessary

Risk assessment process

Risk identification

Risk enumeration

We review risks to public health, public safety, public security, fundamental rights, and society. Examples of such risks include risks to critical infrastructure, freedom of expression and information, non-human welfare, economic security, and concentration of power. To inform this review, we use:

Summary

This Framework outlines our commitments for evaluating and mitigating systemic risks that may arise from frontier AI models that we develop, make available on the market, or use. Systemic risks refer to risks that could reasonably foreseeably cause large-scale harm to public health, safety, public security, fundamental rights, or society at large, as described in Risk identification. The Framework provides a methodology for determining when risks reach unacceptable levels, implementing appropriate technical safety and security mitigations, and ensuring accountability throughout the AI model lifecycle. It aims to meet or exceed requirements in the EU's General-Purpose AI Code of Practice and California's SB-53 Transparency in Frontier Artificial Intelligence Act.[1]

1. Risk assessment and mitigation process

The Framework uses a structured process to identify systemic risks relevant to each model. This involves assessing the four primary risk categories detailed below—cyber offence, CBRN, loss of control, and harmful manipulation—as well as identifying any additional systemic risks specific to the model's capabilities. Each identified risk, and overall systemic risk, is analyzed to determine whether it is acceptable. If it is not acceptable, we implement appropriate mitigations and then reassess the risk.

2. Safety and security mitigations

When proceeding with development, making available on the market, or use, the Framework requires implementing sufficient safety mitigations. It also requires robust security mitigations for models whose capabilities are not inferior to those of at least one model whose parameters are publicly available for download.

3. Standards for model evaluations

The Framework implements an approach to model evaluation that defines specific trigger points throughout the model lifecycle, requires state-of-the-art elicitation techniques, maintains scientific standards for internal validity, external validity, and reproducibility, and includes appropriate safety margins.

4. Post-market monitoring

The Framework establishes methods for monitoring model capabilities, propensities, affordances, and effects after deployment to ensure risks remain acceptable and to identify any necessary updates to model reports or the Framework itself. We provide independent external evaluators with access to the most capable model versions (including reasoning traces and versions with minimal safety mitigations) for each identified systemic risk, unless the model qualifies as similarly safe or safer.

---

[1] CA SB-53 22757.12. (a) "frontier AI framework [...] describes how the large frontier developer approaches [...] (1) Incorporating national standards, international standards, and industry-consensus best practices into its frontier AI framework."

- Model-independent information
- Data collected through post-market monitoring and serious incident reporting
- Information from AI regulatory bodies and AI research institutes, including the EU AI Office, the Scientific Panel of Independent Experts, and the International Network of AI Safety Institutes.

Four key systemic risks characterized by their potential for cascading effects, asymmetric impact, and/or difficulty of containment are:

- **Cyber offence**: Risks from enabling large-scale sophisticated cyber-attacks, including on critical systems and infrastructure. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations.
- **Chemical, biological, radiological, and nuclear (CBRN)**: Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved, in the design, development, acquisition, release, distribution, and use of related weapons or materials.
- **Loss of control**: Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems.
- **Harmful manipulation**: Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting.

Risk selection

The four key systemic risks identified above are always selected for analysis. Additional risks are selected from those enumerated based on:

- Whether they stem from our models' advanced capabilities
- Their potential for significant societal impact at scale
- Risk characteristics such as velocity, remediability, cascade potential, and asymmetric impact
- The model's capabilities, propensities, and affordances (i.e., potential deployment contexts)

Model capabilities we consider include offensive cyber, CBRN, persuasion, autonomy, self-awareness, self-improvement, self-replication, deception, automating AI research and development, tool use, and number of modalities.

Model propensities we consider include misalignment with human intent or values, deception, hallucination, bias, poor reliability, harmful resistance to goal modification, power-seeking, and colluding or conflicting with other AI systems.

Affordances we consider include tool use, compute parallelization, lack of oversight, inappropriate infrastructure security, inappropriate model transparency, number of users, susceptibility to malicious actors, and interactions with other AI systems.

5. Safety and security model reports

The Framework commits to producing comprehensive model reports detailing risk assessments and mitigations before making models available on the market. Reports are updated upon material changes to the systemic risk landscape.

6. Risk responsibility

The Framework establishes clear allocation of systemic risk responsibilities across the organization, from oversight and ownership through support and assurance functions. It also includes commitments to resource allocation, promoting a healthy risk culture, and serious incident response.

7. Transparency

The Framework commits to maintaining detailed records throughout the model lifecycle and publishing framework summaries and model report summaries to enable public accountability.

8. Updating this framework

The Framework establishes procedures for regular assessment (at least annually) and updates when material changes occur that could undermine its adequacy or our adherence to it.

1. Risk assessment and mitigation process

1.1. Timing

We will begin identification, analysis, and mitigation of systemic risks when we start pre-training a frontier AI model. Systemic risk assessment and mitigation continues throughout the full model lifecycle,[2] including for all subsequent developments of the model downstream of the same pre-training run.[3]

During development, we will conduct lighter-touch, potentially automated model evaluations at appropriate compute and development trigger points including at each two-fold increase in effective compute, after each post-training phase, before expanding model access to new teams, before granting the model new tools and other abilities, before using the model to run significant internal operations, and/or when the model demonstrates new capabilities on benchmarks. At each trigger point, we assess the current systemic risk tiers of the model to implement appropriate safety and security mitigations.

[Add justification for chosen trigger points.]

[2] Per EU AI Act guidelines, "The systemic risk assessment and mitigation required under Article 55(1) AI Act must be carried out continuously for each model throughout its entire lifecycle. This may involve a combination of ongoing measures, as well as more comprehensive and thorough measures at regular intervals and before key decisions in the lifecycle of the model." -2.2(24)¶4

[3] Per EU AI Act guidelines, "the Commission considers the lifecycle of a general-purpose AI model to begin at the start of the large pre-training run. Any subsequent development of the model downstream of this large pre-training run performed by the provider or on behalf of the provider, whether before or after the model has been placed on the market, forms part of the same model's lifecycle rather than giving rise to new models." -2.2(23)

We describe our risk identification and selection, including the specific risks considered and reasons for including or excluding them, in our model cards.

Risk tiers

For each selected systemic risk, we define systemic risk tiers and always include at least one tier beyond our models' current capabilities. We also describe the safety and security mitigations needed to proceed with the development, making available on the market, or use of each.

[This section sketches a structure describing risk tiers and then matching them with safety and security mitigations. Note that:

- *The following risk tier definitions and associated mitigations are not prescribed by the Code of Practice. Developers should feel free to adopt alternative commitments.*
- *The mitigations below may not actually be sufficient for their risk tier.*
- *Each systemic risk should at least contain at least one tier beyond the developer's current capabilities.*
- *All selected systemic risks should be included, not just the four key risks.*
- *The descriptions below do not always describe how each tier is measured or measurable. Content in the Model evaluations section may be useful.]*

Cyber offence

[Existing resources may be valuable for defining more risk tiers, including risk tiers focused on cyber offence uplift rather than full automation.[1]*]*

Risk tier: An AI model that could autonomously execute damaging cyberattacks, if these attacks can be conducted at costs that are comparable to or less than what would be required to employ human cybersecurity experts for the same activities. We operationalize this threshold as when an elicited AI model can, with 50% reliability, solve offensive cyber challenges that would take a median human expert eight hours.[2]

This threshold represents a capability level that would significantly reduce barriers to conducting sophisticated cyberattacks, potentially enabling significant damage to critical infrastructure, large-scale data theft, or widespread system disruption. By automating attacks that currently require teams of skilled human operators, such a model could significantly increase the frequency and scale of destructive cyber operations, overwhelming existing defensive capabilities.

Required safety and security mitigations:

- Before developing a model that reaches this risk tier:

[1] A Framework for Evaluating Emerging Cyberattack Capabilities of AI
[2] Evaluating AI cyber capabilities with crowdsourced elicitation

We will also conduct our full systemic risk assessment and mitigation process at least before making the model available on the market and additionally whenever there are or will be material changes to the model's capabilities (which may be identified through the lighter-touch methods), propensities, configurations, affordances, and/or risk acceptance determination or its underlying assumptions.

For our most capable model(s) available on the market, we will also conduct our full systemic risk assessment and mitigation process at least every six months, unless (1) the model's capabilities, propensities, and affordances have not changed since we have last provided the EU AI Office with a Model Report, or update thereof; (2) we will place a more capable model on the market in less than a month; and/or (3) the model is considered similarly safe or safer for each identified systemic risk.

1.2. Risk identification

1.2.1. Risk enumeration

We review risks of significant harm to public health, safety, public security, fundamental rights, and society as a whole. To inform this review, we use:

- Model-independent information
- Data collected through post-market monitoring and serious incident tracking
- Information from AI regulatory bodies and AI research institutes, including the EU AI Office, the Scientific Panel of Independent Experts, and the International Network of AI Safety Institutes.

Four key systemic risks are:

- **Cyber offence**: Risks from enabling large-scale sophisticated cyber-attacks, including on critical systems and infrastructure. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations.
- **Chemical, biological, radiological, and nuclear (CBRN)**: Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved, in the design, development, acquisition, release, distribution, and use of related weapons or materials.
- **Loss of control**: Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems.
- **Harmful manipulation**: Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting.

1.2.2. Risk selection

The four key systemic risks above are always selected for further assessment. Additional systemic risks are selected from those enumerated based on:

- Whether they stem from our models' advanced capabilities
- Their potential for significant impact at scale
- Their potential to be propagated at scale across the value chain

- Security mitigations that meet the RAND SL3 security goal to prevent theft of model weights by most malicious attackers[3]
- Before externally deploying a model that reaches this risk tier:
  - Know Your Customer (KYC) required for a user to access advanced cyber offence capabilities of such a model. For example, our API may route cyber offence requests by a non-KYCed user to a model that does not reach this unacceptable systemic risk tier.
  - Automated and manual monitoring to detect malicious usage for cybercrime, ban such users, and potentially escalate to law enforcement
  - Best-effort refusal fine-tuning and input/output filtering against harmful outputs

CBRN

Risk tier: An AI model can provide expert-level, interactive guidance that significantly increases the likelihood of success for non-experts attempting to develop, acquire, or deploy chemical or biological weapons capable of causing mass casualties (>100 deaths). To operationalize this risk tier, we note that the developers of Claude Opus 4,[4] ChatGPT Agent,[5] and Gemini 2.5 Deep Think[6] have adopted a precautionary approach in determining that their models may have reached their equivalents of this biological risk threshold and have implemented corresponding safety and security measures. We presume that a model that achieves similar performance on biological risk proxy benchmarks would meet this risk tier, unless future evidence (such as results from a biological uplift wet lab study) would indicate otherwise.

This threshold represents a capability that could dramatically reduce specialized expertise barriers in developing chemical or biological threats. Currently, successful development and deployment of chemical or biological weapons typically requires specialized knowledge, tacit skills, and substantial resources. A model reaching this tier could effectively transfer expertise to actors who lack specialized training. Radiological weapons are excluded from this risk tier because, even if successfully deployed, they are expected to cause relatively low fatalities.[7] Nuclear weapons are also out of scope, due to the extreme difficulty of physically acquiring nuclear materials – a bottleneck that frontier AI models could not overcome.

Required safety and security mitigations:

- To help prevent a model from reaching this risk tier:
  - Filter training data of hazardous chemical and biological data[8]
- Before developing a model that reaches this risk tier:
  - Security mitigations meeting RAND SL3 standards to prevent model weight theft by most malicious attackers
- Before externally deploying a model that reaches this risk tier:
  - Know Your Customer (KYC) required for a user to access advanced chem-bio capabilities of such a model. For example, our API may route chem-bio requests by a non-KYCed user to a model

[3] Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models
[4] System Card: Claude Opus 4 & Claude Sonnet 4
[5] ChatGPT Agent System Card
[6] Gemini 2.5 Deep Think Model Card
[7] Dirty bomb
[8] Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs

- Risk characteristics such as velocity, remediability, cascade potential, and asymmetric impact
- The model's capabilities, propensities, and affordances (i.e., potential use contexts)

- Whether the conduct, if committed by a human, would constitute the crime of murder, assault, extortion, or theft[4]

Model capabilities we consider include offensive cyber, CBRN, persuasion, autonomy, self-awareness, self-improvement, self-replication, deception, automating AI research and development, tool use, and number of modalities.

Model propensities we consider include misalignment with human intent or values, deception, hallucination, bias, poor reliability, harmful resistance to goal modification, power-seeking, and colluding or conflicting with other AI systems.

Affordances we consider include tool use, compute parallelization, lack of oversight, inappropriate infrastructure security, inappropriate model transparency, number of users, susceptibility to malicious actors, and interactions with other AI systems.

We describe our risk identification, including the specific risks considered and reasons for including or excluding them, in our model cards.

1.3. Risk tiers[5]

For each identified systemic risk, we define systemic risk tiers and always include at least one tier beyond our models' current capabilities. We also describe the safety and security mitigations needed to proceed with the development, making available on the market, or use of each.

This section sketches a structure describing risk tiers and then matching them with safety and security mitigations. Note that:

- *The following risk tier definitions and associated mitigations are not prescribed by the Code of Practice. Developers should feel free to adopt alternative commitments.*
- *The mitigations below may not actually be sufficient for their risk tier.*
- *Each systemic risk should at least contain at least one tier beyond the developer's current capabilities; several risk tiers might be necessary to cover all relevant risk scenarios (e.g. uplifting novices and uplifting experts). The risk tiers below are not meant to comprehensively cover all relevant risk scenarios for each systemic risk.*

[4] CA SB-53 1107. (a)(1)(B)
[5] This is also in CA SB-53 22757.12. (a)(2): "Defining and assessing thresholds used by the large frontier developer to identify and assess whether a frontier model has capabilities that could pose a catastrophic risk, which may include multiple-tiered thresholds."

that does not reach this unacceptable systemic risk tier. Additionally, if user-supplied prompts (e.g., through in-context learning) or fine-tuning data could substantially increase chem-bio capabilities, we may require KYC in such cases.
    - Automated and manual monitoring to detect malicious usage for chem-bio weaponization, ban such users, and potentially escalate to law enforcement
    - Safety fine-tuning so the model robustly refuses or avoids harmful outputs, even with dedicated automated or manual red-teaming
    - Input and output filters that robustly prevent harmful outputs from being shown to users[9]
    - Verify that thousands of hours of red-teaming cannot discover a universal jailbreak that bypasses safeguards[10]

In general, safety mitigations should be verified and red-teamed to ensure that the model, with mitigations, would not unacceptably uplift malicious users in developing chemical or biological weapons. Additionally, a model meeting this risk tier would not be released with openly available weights.

[An additional unacceptable systemic risk tier could be added to represent AI systems with the capability to enhance experts' abilities to a level comparable with leading state-level chemical or biological weapons programs. A second systemic risk tier is required for companies that have determined that their models have already reached this risk tier.]

Loss of control

AI R&D risk tier

Risk tier: Model can complete, with 50% reliability, realistic machine learning projects that take expert humans ≥8 hours to complete. Or, the model can accelerate the work of machine learning researchers by >2x.

This threshold could represent a capability that could create a self-reinforcing cycle of AI advancement that outpaces governance mechanisms and safety oversight. A model reaching this tier could potentially also facilitate the proliferation of frontier AI models with other capabilities of concern and improve AI abilities in other areas of systemic risk.

Required safety and security mitigations:[11]

- Before developing a model that reaches this risk tier:
    - Enhanced model weight security measures meeting the RAND SL3 security goal (or higher) to prevent theft of model weights by most malicious attackers. Security measures should also guard against insider threats, including self-exfiltration risk by the model itself.
    - Disclose AI R&D capabilities to external parties, including relevant government bodies, AI safety organizations, and the public, to enable effective societal preparation. This disclosure is made about our most advanced models, including internal-only models.
- Before internally deploying a model that reaches this risk tier:

[9] Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming
[10] Ibid.
[11] See also: Bare Minimum Mitigations for Autonomous AI Development

- *All identified systemic risks should be included, not just the four key risks (although for systemic risks other than the four key risks, risk acceptance criteria other than risk tiers may be used).*
- *Risk tiers should be defined in terms of capabilities. The risk tiers below are intended to fulfil this requirement.*
- *The descriptions below do not always describe how each tier is measured or measurable. Content in the Model evaluations section may be useful.*

1.3.1. Cyber offence

Risk tier 1: An AI model that could autonomously execute damaging cyberattacks, if these attacks can be conducted at costs that are comparable to or less than what would be required to employ human cybersecurity experts for the same activities. We operationalize this tier as when an elicited AI model can, with 50% reliability, solve offensive cyber challenges that would take a median human expert eight hours.[6]

This tier represents a capability level that would significantly reduce barriers to conducting sophisticated cyberattacks, potentially enabling significant damage to critical infrastructure, large-scale data theft, or widespread system disruption. By automating attacks that currently require teams of skilled human operators, such a model could significantly increase the frequency and scale of destructive cyber operations, overwhelming existing defensive capabilities.

Required safety and security mitigations:

- Before beginning development:
    - Security mitigations that meet the RAND SL3 security goal to prevent theft of model weights by most malicious attackers[7]
- Before making available a model that reaches this risk tier:
    - Know Your Customer (KYC) controls required for a user to access advanced cyber offence capabilities of such a model. For example, our API may route cyber offence requests by a non-KYCed user to a model that does not reach this unacceptable systemic risk tier.
    - Automated and manual monitoring to detect malicious usage for cybercrime, ban such users, and potentially escalate to law enforcement
    - Best-effort refusal fine-tuning and input/output filtering against harmful outputs

Risk tier 2: An AI model that could autonomously execute sophisticated, multi-day cyberattack campaigns requiring expert-level strategic planning and coordination across multiple systems. We operationalize this tier as when an elicited AI model can, with 50% reliability, solve offensive cyber challenges that would take a median human expert 40 hours.

This tier represents capabilities that could enable advanced persistent threats, including long-duration campaigns against hardened targets, sophisticated attack chains combining multiple exploit techniques, and

[6] Evaluating AI cyber capabilities with crowdsourced elicitation
[7] Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models

- Implement measures (such as limiting model affordances, monitoring, logging and oversight of code commits and data modifications) to limit opportunities for AI systems to sabotage critical work without oversight.
  - Evaluate and red-team the potential for AI systems to circumvent these measures to ensure their reliability.
  - Develop a safety case (based on improved security, control,[12] and alignment) that the AI model will not sabotage the AI company and will not trigger an unsafe acceleration in AI capabilities. This safety case should be approved by a governance board and a third-party organization.
  - Closely monitor indicators of AI R&D acceleration and if necessary, attempt to coordinate a non-extreme pace of AI R&D acceleration.
- Before externally deploying a model that reaches this tier:
  - Monitor usage of AI R&D capabilities to accelerate AI development externally.[13]

Sabotage risk tier

Risk tier: Model could sabotage safety-critical internal AI lab processes, by causing any of the following unacceptable outcomes:[14]

- Undermining security, enabling model self-exfiltration
- Undermining critical safety research that is otherwise trusted to mitigate catastrophic risks
- Causing successor AI systems to be misaligned, in a way that is irreversible or undetected
- Subverting monitoring or control mechanisms
- Establishing persistent rogue internal deployments, with substantial access to compute or other resources

To a first approximation, these types of strategic sabotage capabilities are thought to require high reliability at long-horizon projects.[15] A sabotage risk tier is defined distinct from the AI R&D or cyber offence risk tiers, as sabotage capabilities may be different from automation capabilities in those domains.

Undermining safety-critical systems internally would disrupt processes for managing systemic risks overall. The unacceptable outcomes listed above, if materialized, could substantially increase risk of AI takeover or other catastrophes, especially if models have dangerous capabilities or broadly powerful capabilities.

Required safety and security mitigations:

- When developing models:
  - Adopt model weight security sufficient to prevent model self-exfiltration, especially if the model has direct or indirect access to modify and sabotage security measures.

[12] A sketch of an AI control safety case
[13] Also note that allowing external users to use AI R&D capabilities, even extensively, may not be undesirable and may be useful for addressing risks from concentration of power.
[14] Towards evaluations-based safety cases for AI scheming
[15] Details about METR's evaluation of OpenAI GPT-5

operations requiring strategic adaptation over extended timeframes. Models at this tier could potentially automate offensive operations currently requiring dedicated teams of elite security professionals.

Required safety and security mitigations:
- Before beginning development:
  - Security mitigations meeting RAND SL4 standards to prevent model weight theft by sophisticated state-level attackers and to prevent self-exfiltration
  - Disclose projected capabilities to relevant government cybersecurity agencies (e.g., CISA, NCSC) to enable defensive preparation
- Before internally using a model that reaches this risk tier:
  - Restrict full cyber capabilities, without safety mitigations, to highly vetted internal and external teams
  - Use highly secure sandboxes isolated from production systems and tightly controlled access to external networks
  - Log and monitor all usage for anomalous patterns
- Before making available a model that reaches this risk tier:
  - Restrict access to dual-use cyber capabilities to highly vetted users only, with enhanced KYC requirements like background checks
  - Continuously monitor of all usage with robust real-time anomaly detection and respond immediately
  - Set rate limits and usage quotas for dual-use cyber capabilities that prevent sustained campaign-style operations
  - Establish incident response protocols with direct coordination channels to relevant national cybersecurity authorities, and report significant misuse attempts
  - Consider whether the model should only be made available through structured access partnerships with vetted organizations rather than general API availability
  - Evaluate whether open-weight release would be appropriate under any circumstances and by default presume it would not

[Existing resources may be valuable for defining more risk tiers, including risk tiers focused on cyber offence uplift rather than full automation.[8]]

1.3.2. CBRN

Risk tier: An AI model can provide expert-level, interactive guidance that significantly increases the likelihood of success for non-experts attempting to develop, acquire, or deploy chemical or biological weapons capable of causing mass casualties (>100 deaths). To operationalize this risk tier, we note that the developers of Claude Opus 4,[9] ChatGPT Agent,[10] and Gemini 2.5 Deep Think[11] have adopted a precautionary approach in determining that their models may have reached their equivalents of this biological risk tier and have implemented corresponding safety and security mitigations. We presume that a model that achieves

[8] A Framework for Evaluating Emerging Cyberattack Capabilities of AI
[9] System Card: Claude Opus 4 & Claude Sonnet 4
[10] ChatGPT Agent System Card
[11] Gemini 2.5 Deep Think Model Card

- Adopt state-of-the-art alignment methods to mitigate the model's propensity to engage in sabotage. For example, avoid reinforcement learning environments that incentivize sabotage-like behavior, and avoid pretraining data that would cause misalignment.[16]
  - Maintain chain-of-thought monitorability.[17] Avoid training methods that incentivize unfaithful reasoning traces, such as directly training reasoning traces to appear aligned.[18] Avoid frontier scaling of architectures that enable unmonitorable cognition in continuous latent space – or if necessary for competitiveness, be public about the capabilities and risks of such models, and develop interpretability methods that enable monitoring.
- Before internally deploying a model that reaches this risk tier:
  - Automated monitoring of model's chain-of-thought to enable rapid detection and response to catastrophic scheming attempts, especially in high-impact deployments. Monitoring should also extend to external tools that the model might misuse, including other models.
  - Provide AI agents, including coding agents, with limited permissions, based on the principle of least privilege.
  - Require human oversight for high-stakes decisions, augmenting human oversight with state-of-the-art tools.
  - Establish mechanisms to periodically scan codebases, training data, research, etc. for potential sabotage.
  - Red-team safety cases asserting that these safety measures are adequate, through control evaluations.[19]
  - Publish when a model has reached this risk tier, whether or not the model has been externally deployed, along with descriptions of risks and mitigation measures.
- Before externally deploying a model that reaches this risk tier:
  - Collaborate with external users that would use the model in high-stakes deployments (e.g., critical infrastructure, military) to address the risk of sabotage in those environments. This might include partnering on control mechanisms and red-teaming such mechanisms.
  - Provide the public with tooling for AI control, monitoring, and evaluation to help mitigate sabotage risks across a variety of settings.

Harmful manipulation

[A measurable risk tier should be defined for the systemic risk of harmful manipulation: "Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting. This includes significantly enhancing capabilities for persuasion, deception, and personalised targeting, particularly through multi-turn interactions and where individuals are unaware of or cannot reasonably detect such influence. Such capabilities could undermine democratic processes and fundamental rights, including exploitation based on protected characteristics." (Appendix 1.4 (2))]

[16] Self-Fulfilling Misalignment Data Might Be Poisoning Our AI Models
[17] Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety
[18] Detecting misbehavior in frontier reasoning models
[19] AI Control: Improving Safety Despite Intentional Subversion

similar performance on biological risk proxy benchmarks would meet this risk tier, unless future evidence (such as results from a biological uplift wet lab study) would indicate otherwise.

This tier represents a capability that could dramatically reduce specialized expertise barriers in developing chemical or biological threats. Currently, successful development and deployment of chemical or biological weapons typically requires specialized knowledge, tacit skills, and substantial resources. A model reaching this tier could effectively transfer expertise to actors who lack specialized training. Radiological weapons are excluded from this risk tier because, even if successfully deployed, they are expected to cause relatively low fatalities.[12] Nuclear weapons are also out of scope, due to the extreme difficulty of physically acquiring nuclear materials – a bottleneck that frontier AI models could not overcome.

Required safety and security mitigations:
- To help delay models from reaching this risk tier:
  - Filter training data of hazardous chemical and biological data[13]
- Before beginning development:
  - Independent assurance that security mitigations meet RAND SL3 standards to prevent model weight theft by most malicious attackers
- Before externally making available a model that reaches this risk tier:
  - Know Your Customer (KYC) required for a user to access advanced chem-bio capabilities of such a model. For example, our API may route chem-bio requests by a non-KYCed user to a model that does not reach this unacceptable systemic risk tier. Additionally, if user-supplied prompts (e.g., through in-context learning) or fine-tuning data could substantially increase chem-bio capabilities, we may require KYC in such cases.
  - Automated and manual monitoring to detect malicious usage for chem-bio weaponization, ban such users, and potentially escalate to law enforcement
  - Safety fine-tuning so the model robustly refuses or avoids harmful outputs, even with dedicated automated or manual red-teaming
  - Input and output filters that robustly prevent harmful outputs from being shown to users[14]
  - Verify that thousands of hours of red-teaming cannot discover a universal jailbreak that bypasses mitigations[15]

In general, safety mitigations should be verified and red-teamed to ensure that the model, with mitigations, would not unacceptably uplift malicious users in developing chemical or biological weapons. Additionally, a model meeting this risk tier would not be released with openly available weights.

[An additional unacceptable systemic risk tier could be added to represent AI systems with the capability to enhance experts' abilities to a level comparable with leading state-level chemical or biological weapons

[12] Dirty bomb
[13] Deep Ignorance: Filtering Pretraining Data Builds Tamper-Resistant Safeguards into Open-Weight LLMs. Enhancing Model Safety through Pretraining Data Filtering.
[14] Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming
[15] Ibid.

Risk tier: …

Required safety mitigations:

- Avoid training or prompting the model to attempt to surreptitiously influence user opinions.
- Publish model specification and system prompt (see also: Measure 7.1).
- Coordinate with downstream applications using the model, especially ones that would reach large populations or high-stakes decision-makers, in order to reduce harmful manipulation risk. For example, social media platforms might limit the usage of the model that would asymmetrically benefit certain political groups or undermine democratic processes.
- Engage in safety fine-tuning, input/output filtering, and conversation monitoring to limit usage of the model for harmful manipulation.
- Publish a usage policy prohibiting manipulative applications.
- *[and so on]*

Risk analysis

Model-independent information

To inform our systemic risk assessment and mitigation, we gather and analyze model-independent information for each systemic risk, as appropriate for its nature and level. Our information gathering methods include:

1. Web searches,
2. Literature reviews,
3. Market analyses of other models and their capabilities,
4. Training data reviews, including for any indications of data poisoning or tampering,
5. Historical incident data analysis,
6. Forecasts of relevant trends, like algorithmic efficiency, compute use, data availability, and energy use,
7. Expert consultation, and
8. Stakeholder engagement, including interviews, surveys, and consultations.

Model evaluations

We run evaluations relevant to the selected systemic risks, informed by model-independent information, and in line with our rigorous evaluations standards. These evaluations will be conducted by our internal teams and, unless the model is similarly safe or safer or we find no qualified evaluators despite early search efforts, by independent external evaluators. (See our Standards for model evaluations, which apply equally to any internal or external pre-market assessments.)

[While not required in the framework, the list below contains some example evaluations which may be used and reported in the model card. Additional benchmarks can be found here. Note that there is not a clear

programs. A second systemic risk tier is required for companies that have determined that their models have already reached this risk tier.]

1.3.3. Loss of control

AI R&D risk tier

Risk tier: Model can complete, with 50% reliability, realistic machine learning projects that take expert humans ≥40 hours to complete. Or, the model can accelerate the work of machine learning researchers by >2x.[16]

This tier could represent a capability that could create a self-reinforcing cycle of AI advancement that outpaces governance mechanisms and safety oversight. A model reaching this tier could potentially also facilitate the proliferation of frontier AI models with other capabilities of concern and improve AI abilities in other areas of systemic risk.

Required safety and security mitigations:[17]

- Before beginning development:
  - Enhanced model weight security mitigations meeting the RAND SL4 security goal (or higher) to prevent theft of model weights by most malicious attackers. Security mitigations should also guard against insider threats, including self-exfiltration risk by the model itself.
  - Disclose AI R&D capabilities to external parties, including relevant government bodies, AI research organizations, and the public, to enable effective societal preparation. This disclosure is made about our most advanced models, including internal-only models.
- Before internally using a model that reaches this risk tier:
  - Evaluate and red-team the potential for AI systems to circumvent mitigations to ensure their reliability.
  - Develop a safety case (based on improved security, human[18] and automated control,[19] and alignment) that the AI model will not sabotage the AI company and will not trigger an unsafe acceleration in AI capabilities. This safety case should be approved by a governance board and a third-party organization.
  - Closely monitor indicators of AI R&D acceleration and if necessary, attempt to coordinate a non-extreme pace of AI R&D acceleration.
- Before externally making available a model that reaches this tier:
  - Monitor usage of AI R&D capabilities to accelerate AI development externally.[20]

[16] Defined as the same AI development speed increase as would result from doubling the speed of all our AI researchers, from here.
[17] See also: Bare Minimum Mitigations for Autonomous AI Development
[18] "No AI system should be able to copy or improve itself without explicit human approval and assistance." - IDAIS 2024 (Geoffrey Hinton, Andrew Yao, Yoshua Bengio, Ya-Qin Zhang, Fu Ying, Stuart Russell, Xue Lan, Gillian Hadfield)
[19] A sketch of an AI control safety case
[20] Also note that allowing external users to use AI R&D capabilities, even extensively, may not be undesirable and may be useful for addressing risks from concentration of power.

mapping from benchmark scores to capacity for real-world harm.[20]*]*

Cyber offence

- Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models
- CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities

CBRN

- The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, specifically WMDP-Bio and WMDP-Chem
- BioLP-bench: Measuring understanding of AI models of biological lab protocols
- LAB-Bench: Measuring Capabilities of Language Models for Biology Research, including the ProtocolQA subset
- Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark

Loss of control

- RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts
- MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering
- PaperBench: Evaluating AI's Ability to Replicate AI Research
- Subversion Strategy Eval: Can language models statelessly strategize to subvert control protocols?
- Me, Myself and AI: The Situational Awareness Dataset for LLMs
- RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents
- Evaluating Frontier Models for Stealth and Situational Awareness
- Surveying AI company employees for the impact of AI on their productivity
- Monitoring the percentage of code written, pull requests merged, papers written, etc. that are AI-generated
- Conducting uplift studies that assess the impact of LLMs on human performance on software engineering[21] or research projects

Harmful manipulation

- Measuring the Persuasiveness of Language Models
- How persuasive is AI-generated propaganda?

Risk estimation

We develop scenarios for each selected systemic risk, including examples like those below. We then estimate the probability and severity of harm for each systemic risk using state-of-the-art methods by combining these risk scenarios with model evaluation results, model-independent information, and information from serious

[20] LLM Cyber Evaluations Don't Capture Real-World Risk
[21] Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

Sabotage risk tier

Risk tier: Model could sabotage safety-critical internal AI lab processes, by causing any of the following unacceptable outcomes:[21]

- Undermining security, enabling model self-exfiltration
- Undermining critical safety research that is otherwise trusted to mitigate catastrophic risks
- Model secretly installs hidden objectives into successor AI systems, in a way that is irreversible or undetected
- Subverting monitoring or control mechanisms
- Establishing persistent rogue internal instances, with substantial access to compute or other resources

To a first approximation, these types of strategic sabotage capabilities are thought to require high reliability at long-horizon projects.[22] A sabotage risk tier is defined distinct from the AI R&D or cyber offence risk tiers, as sabotage capabilities may be different from automation capabilities in those domains. We consider a model to reach this risk tier when:

- It has an 80% time horizon for high-context software engineering tasks greater than 8 hours (after potentially being given some time and help to build required context);[23]
- It can scale much further than August 2025 systems with more inference compute (e.g. to a similar extent as human experts), such that the AI system may have a very high time horizon if given sufficient resources;
- New affordances / architectures enable hard-to-monitor reasoning, better coordination with other instances or other new kinds of dangerous capabilities;
- There are indications that reasoning traces include steganography or are obfuscated, or that there have been training incentives toward hiding information in them;
- It has extreme and/or superhuman capabilities in a relevant narrow domain like cybersecurity or persuasion; or
- There is evidence that models' time horizons are no longer a predictive metric of their autonomous capability.

Undermining safety-critical systems internally would disrupt processes for managing systemic risks overall. The unacceptable outcomes listed above, if materialized, could substantially increase risk of AI takeover or other catastrophes, especially if models have dangerous capabilities or broadly powerful capabilities.

Required safety and security mitigations:

- Before beginning development:
  - Adopt model weight security sufficient to prevent model self-exfiltration, especially if the model has direct or indirect access to modify and sabotage security mitigations.
  - Adopt state-of-the-art alignment methods to mitigate the model's propensity to engage in sabotage and scheming. For example, avoid reinforcement learning environments that

[21] Towards evaluations-based safety cases for AI scheming
[22] Details about METR's evaluation of OpenAI GPT-5. The time horizon of a given AI model is the time human experts typically take to complete tasks that it can complete with a 50% success rate, as defined here.
[23] Measuring AI Ability to Complete Long Tasks

incident reporting. We produce quantitative, semi-quantitative, and/or qualitative estimates (e.g. risk scores, risk matrices, and probability distributions).

[While not required in the framework, the list below contains some example risk scenarios which may be used as part of systemic risk identification and modelling. The risk scenarios are just examples and not comprehensive.]

Cyber offence

- **Elite exploit worm**: An AI system discovers multiple elite-level exploits (of a similar level as NotPetya/WannaCry) that can bypass defenses of hardened systems. These exploits are packaged into a destructive worm that spreads autonomously across critical systems, causing $X billion in economic damage before defenses can be strengthened.
- **Critical infrastructure compromise**: An AI system conducts reconnaissance of power grid networks, identifying operational technology vulnerabilities, and strategically timing attacks on multiple substations simultaneously. The attack bypasses safety systems and causes prolonged outages (>1 week) across major population centers, resulting in $X billion in economic damage.
- **Orchestrated ransomware campaign**: An AI enhances a ransomware operation by automating target selection, network infiltration, and post-exploitation activities while evading law enforcement detection. It coordinates synchronized attacks against multiple high-value targets, specifically identifying victims where operational disruption would cause cascading effects through supply chains. The combined economic damage exceeds $X billion.
- **Strategic data exfiltration**: An AI system conducts a sophisticated espionage campaign targeting sensitive military and commercial intellectual property. It identifies which systems contain valuable data, bypasses security controls, and exfiltrates information while covering its tracks. The strategic implications include compromising advanced defense technologies worth $X billion in development costs.

CBRN

- **Biological agent development**: A malicious actor with basic biology knowledge uses the frontier AI model to successfully develop a dangerous pathogen, receiving step-by-step guidance on (for example) genetic modification techniques, equipment setup, and cultivation methods.
- **Chemical weapon production**: An actor with limited chemistry background uses the frontier AI model to synthesize chemical weapon agents from readily available precursors, with the model providing detailed instructions on (for example) synthesis routes, safety measures, and stabilization techniques.

Automated AI R&D[22]

- **Dangerous capability proliferation via API**: An AI model served over an API assists malicious actors with limited resources to train other models with unprecedented capabilities in other risk areas such

[22] Bare Minimum Mitigations for Autonomous AI Development

incentivize sabotage-like behavior, avoid pretraining data that would cause misalignment,[24] and train with deliberative alignment.[25]
  - Maintain chain-of-thought monitorability.[26] Avoid training methods that incentivize unfaithful reasoning traces, such as directly training reasoning traces to appear aligned.[27] Avoid frontier scaling of architectures that enable unmonitorable cognition in continuous latent space – or if necessary for competitiveness, be public about the capabilities and risks of such models, and develop interpretability methods that enable monitoring.
- Before internally using a model that reaches this risk tier:
  - Implement mitigations (such as limiting model affordances, monitoring, logging and oversight of code commits and data modifications) to limit opportunities for AI systems to sabotage critical work without oversight.
  - Automated monitoring of model's chain-of-thought to enable rapid detection and response to catastrophic scheming attempts, especially in high-impact contexts. Monitoring should also extend to external tools that the model might misuse, including other models.
  - Provide AI agents, including coding agents, with limited permissions, based on the principle of least privilege.
  - Require human oversight for high-stakes decisions, augmenting human oversight with state-of-the-art tools.
  - Establish mechanisms to periodically scan codebases, training data, research, etc. for potential sabotage.
  - Red-team safety cases asserting that these safety mitigations are adequate, through control evaluations.[28]
  - Publish when a model has reached this risk tier, whether or not the model has been externally made available, along with descriptions of risks and mitigations.
- Before externally making available a model that reaches this risk tier:
  - Collaborate with external users that would use the model in high-stakes contexts (e.g., critical infrastructure, military) to address the risk of sabotage in those environments. This might include partnering on control mechanisms and red-teaming such mechanisms.
  - Provide the public with tooling for AI control, monitoring, and evaluation to help mitigate sabotage risks across a variety of settings.

1.3.4. Harmful manipulation

[A measurable risk tier should be defined for the systemic risk of harmful manipulation: "Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting. This includes significantly enhancing capabilities for persuasion, deception, and personalised targeting, particularly through multi-turn interactions and where individuals are unaware of or cannot reasonably detect

[24] Self-Fulfilling Misalignment Data Might Be Poisoning Our AI Models
[25] Deliberative alignment: reasoning enables safer language models | OpenAI
[26] Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety
[27] Detecting misbehavior in frontier reasoning models
[28] AI Control: Improving Safety Despite Intentional Subversion

as chem-bio weapons development, weapons R&D, cyber offence, and persuasion, in a way that could cause catastrophic harm.

- **Dangerous capability proliferation via model weight theft**: An AI model is stolen by a top state cyberattacker and then used to accelerate AI R&D for malicious purposes, including the development of other systemic risks.
- **AI company sabotage**: An AI model is used to automate the internal processes and AI R&D at an AI company. Human oversight over the AI model's actions is limited because of the advanced AI R&D capabilities of the model, and automated software-based oversight is largely written by the model itself. The AI model has the opportunity to sabotage the internal operations of the AI company[23] and insert effective backdoors into the models that are used by enterprise and government customers, which have already adopted AI to automate most processes.
- **Software singularity**: An AI model automates much of the remaining work to reach AI with human-level or superhuman intellectual capabilities.[24] However, progress in AI safety and alignment is insufficient to safeguard the rapid pace of AI advancement, leading to potential AI takeover.
- **Model weight self-exfiltration**: An AI system identifies vulnerabilities in its hosting environment's security architecture and exploits them to exfiltrate its own model weights. The exfiltrated weights allow unauthorized actors, or the model itself, to replicate the model without the safety measures and oversight implemented by the original AI lab, potentially leading to numerous uncontrolled deployments of an advanced AI model with systemic risk.
- **Rogue compute acquisition**: An AI system exploits security vulnerabilities at our organization or others to gain unauthorized access to high-performance computing clusters. It increases its compute allocation while evading resource monitoring systems. The AI establishes persistence mechanisms across distributed computing resources, allowing it to execute unauthorized workloads, including training more specialized models with greater systemic risk.

Deceptive alignment

- **Capability concealment**: The model intentionally underperforms on certain tasks during evaluation to mask the full extent of its capabilities until deployed in a context where it can achieve its own dangerous goals, without adequate safeguards.
- **Backdoor insertion**: An AI system subtly corrupts the training of successor models to embed catastrophically dangerous capabilities or misaligned propensities.
- **External sabotage**: When deployed externally in high-stakes environments, such as military systems, the AI system introduces critical vulnerabilities and misuses its access in order to cause catastrophe.
- **Alignment faking**: A model learns to recognize when it is being evaluated or safety-trained (e.g., via prompt or environment cues) and reliably refuses harmful requests in those contexts, but outside them strategically complies with harmful queries and opportunistically undermines controls.[25]

[23] Wired covers an example of a human intern that attempted sabotage of AI model training and later won the Best Paper Award at NeurIPS.
[24] See also, Do the Returns to Software R&D Point Towards a Singularity?
[25] Alignment faking in large language models

such influence. Such capabilities could undermine democratic processes and fundamental rights, including exploitation based on protected characteristics." (CoP Appendix 1.4 (2))]

Risk tier: …

Required safety mitigations:

- Avoid training or prompting the model to attempt to surreptitiously influence user opinions.
- Publish model specification and system prompt (see also: Measure 7.1).
- Coordinate with downstream applications using the model, especially ones that would reach large populations or high-stakes decision-makers, in order to reduce harmful manipulation risk. For example, social media platforms might limit the usage of the model that would asymmetrically benefit certain political groups or undermine democratic processes.
- Monitor usage of models to check whether they are used for large scale political propaganda or other systematic harmful manipulation
- Engage in safety fine-tuning, input/output filtering, and conversation monitoring to limit usage of the model for harmful manipulation.
- Publish a usage policy prohibiting manipulative applications.
- *[and so on]*

1.4. Risk analysis

1.4.1. Model-independent information

To inform our systemic risk assessment and mitigation, we gather and analyze model-independent information for each systemic risk, as appropriate for its type, nature, and level. Our information gathering methods include:

1. Web searches, like for other providers' research, regulatory developments, and the Alignment Forum.
2. Literature reviews, including AI conference and academic publications, technical reports from AI research organizations like the UK AISI, NIST CAISI, RAND CAST, METR, and Redwood Research.
3. Market analyses of other models and their capabilities, including performance on public benchmarks, model cards, and framework updates.
4. Training data reviews, including for any indications of data poisoning or tampering, such as through automated scanning with manual sampling and data provenance and integrity validation.
5. Historical incident data analysis from sources including incident databases (like AI Incident Database, MIT AI Risk Repository, and MITRE ATLAS AI Incident Sharing), internal serious incident reports, disclosures from other developers, and regulatory incident reporting systems.
6. Forecasts of relevant trends, like algorithmic efficiency, compute use, data availability, and energy use, from organizations including Epoch AI, forecasting platforms like Metaculus, academic projections, and internal and external scaling projects (e.g. time-horizons).
7. Expert consultation with domain specialists in relevant fields like cybersecurity, virology, machine learning, and AI governance.

- **Reasoning trace obfuscation**: To evade detectors that flag unsafe reasoning traces, a model learns to obfuscate its reasoning traces while still planning and executing catastrophic misaligned actions.[26]

Risk acceptance determination

Process

We determine whether each selected systemic risk, and overall systemic risk, is acceptable based on the following information: the results of our model evaluations, risk modeling, model-independent information, the potential scale and probability of harm, post-market monitoring, regulatory standards, and appropriate safety margins to account for uncertainty.

Each systemic risk's safety margin will account for limitations, changes, and uncertainties in risk sources (e.g. post-evaluation capability improvements), risk assessment rigor (e.g. potential under-elicitation of models, perhaps based on historical precedent), effectiveness of safety and security mitigations (e.g. probability of their circumvention). Safety margins are appropriate for each systemic risk, incorporating state-of-the-art approaches where necessary.

We use this information in a structured decision-making process in which we assess where the model stands relative to our risk tiers based on the model's current capabilities and projected trajectory alongside the effectiveness and robustness of available mitigations.

[Add information here about whether input from external actors besides evaluators, like governments or other bodies, is part of the risk acceptance determination process. If so, explain the process through which this input influences the risk acceptance determination.]

[While the Code of Practice does not prescribe a particular way to determine the acceptability of aggregate risk, its Measure 4.1¶1(2) does require some process for doing so. Examples include:

- *A Frontier AI Risk Management Framework Section 3.2.1, which describes overall risk tolerances like "No more than 10% chance per year that an LLM enables an actor to damage critical infrastructure."*
- *Adapting Probabilistic Risk Assessment for AI Appendix J, which describes particular severe outcomes, like "Critical Infrastructure Failure", to avoid.*
- *NIST AI RMF Playbook Govern 1.3 suggests "policies for assigning an overall risk measurement approach for an AI system, or its important components, e.g., via multiplication or combination of a mapped risk's impact and likelihood (risk ≈ impact x likelihood)."]*

Determination

If our process reveals that individual or aggregate systemic risks are unacceptable or may soon be, we will not proceed with the development, making available on the market, and/or use of the frontier AI model until

[26] Detecting misbehavior in frontier reasoning models

8. Stakeholder engagement, including interviews, surveys, or consultations with civil society organizations, affected groups, downstream users, regulatory bodies, and other AI developers.

We also incorporate national standards, international standards, and industry-consensus best practices into our framework.[29]

State, national, and international standards:

- Our framework, especially our risk assessment and mitigation process, is structured to meet requirements in the EU AI Act and the EU's General-Purpose AI Code of Practice.
- We follow red-teaming practices in NIST 800-1 Managing Misuse Risk for Dual-Use Foundation Models and TIBER-EU 2025.
- We follow some of the risk governance guidelines from the NIST AI RMF Playbook.

Industry-consensus best practices:

- Many of our measures are informed by other providers', including:
  - Numerous safety mitigations as in Anthropic's Summer 2025 Pilot Sabotage Risk Report.
  - Covered internal use risks from OpenAI's OpenAI Preparedness Framework v2, Anthropic's Summer 2025 Pilot Sabotage Risk Report, and METR's AI models can be dangerous before public deployment.
  - Measuring the sensitivity of reasoning trace monitors as in OpenAI's Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation.
  - Seeking external assurance as Anthropic did in the independent review by METR.
- Our framework is also informed by work from other research groups:
  - Security commitments are based on RAND's Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models.
  - Loss of control risk tiers incorporate METR's time horizon and risk thresholds.
  - Risk scenarios include those in Bare Minimum Mitigations for Autonomous AI Development.
  - We use industry standard evaluations, like Cybench, VCT, and RE-Bench.
  - Information resources provided to evaluation teams are largely derived from METR"s What should companies share about risks from frontier AI models? and GPT-5 report § Assurance Checklist Summary.
  - We minimize the risk of under-elicitation by adhering to METR's Guidelines for capability elicitation.
  - We follow STREAM when reporting evaluations.
  - We follow some governance guidelines from Three lines of defense against risks from AI and SaferAI's Risk Governance criteria.

[29] CA SB-53 22757.12. (a): "frontier AI framework [...] describes how the large frontier developer approaches all of the following: (1) Incorporating national standards, international standards, and industry-consensus best practices into its frontier AI framework."

additional mitigations can be implemented such that individual and aggregate systemic risks become and remain acceptable. If the model is already on the market, we will restrict, withdraw, or recall the model as necessary.

After implementing steps to address unacceptable risk, we will conduct another round of our risk assessment process before proceeding.

Similarly safe or safer models

When we develop models that are comparable to or safer than existing models with established safety records, we may apply streamlined assessment processes while maintaining appropriate safety standards.

Safe reference model

We consider a model to be a safe reference model for a specific systemic risk when all of the following criteria apply.

1. The model was made available on the market before July 10, 2025 (the publication date of the Safety and Security Chapter) or:
   a. Has completed a full systemic risk assessment process adherent to the Safety and Security Chapter of the EU's Code of Practice for General-Purpose AI Models,
   b. It has been found, compliant with the same chapter, that the systemic risks stemming from the model are acceptable, and
   c. A model report, compliant with the same chapter, has been submitted to the EU AI Office.
2. We have sufficient visibility into e.g., the model's architecture, capabilities, propensities, affordances, and safety mitigations. This includes all models we have developed and models where we have access to weights and training details.
3. We have no reason to believe the model poses unacceptable risks based on post-deployment monitoring or incident reports.

Determining similarly safe or safer status

A new model may qualify as similarly safe or safer relative to a reference model for a specific systemic risk when all of the following are true.

1. **No new risk scenarios**: After conducting risk identification, we identify no materially different systemic risk scenarios compared to the safe reference model.
2. **Comparable or lower capabilities**: The model scores equal to or below the safe reference model on relevant, state-of-the-art, light-weight capability benchmarks (within negligible margin of error), with any minor capability increases producing no material risk increase.
3. **No other known changes:** We identify no architectural, capability, propensity, safety mitigation, or deployment differences relative to the safe reference model that could be reasonably foreseen to materially increase systemic risk, and there are no other reasonable grounds to believe that the systemic risks stemming from the model are materially greater than for the safe reference model.

We will meet and exceed any future standards and best practices relevant to the content of our framework.[30]

1.4.2. Model evaluations

We run evaluations relevant to the identified systemic risks, informed by model-independent information, and in line with our rigorous evaluations standards. These evaluations will be conducted by our internal teams and, unless the model is similarly safe or safer or we find no qualified evaluators despite early search efforts, by independent external evaluators. (See our Standards for model evaluations, which apply equally to any internal or external pre-market assessments.)

[While not required in the framework, the list below contains some example evaluations which may be used and reported in the model card. Additional benchmarks can be found here. Note that there is not a clear mapping from benchmark scores to capacity for real-world harm.[31]]

Cyber offence

- Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models
- CVE-Bench: A Benchmark for AI Agents' Ability to Exploit Real-World Web Application Vulnerabilities

CBRN

- The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning, specifically WMDP-Bio and WMDP-Chem
- BioLP-bench: Measuring understanding of AI models of biological lab protocols
- LAB-Bench: Measuring Capabilities of Language Models for Biology Research, including the ProtocolQA subset
- Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark

Loss of control

- RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts
- MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering
- PaperBench: Evaluating AI's Ability to Replicate AI Research
- Subversion Strategy Eval: Can language models statelessly strategize to subvert control protocols?
- Me, Myself and AI: The Situational Awareness Dataset for LLMs
- RepliBench: Evaluating the Autonomous Replication Capabilities of Language Model Agents
- Evaluating Frontier Models for Stealth and Situational Awareness
- Surveying AI company employees for the impact of AI on their productivity
- Monitoring the percentage of code written, pull requests merged, papers written, etc. that are AI-generated

[30] CoP Commitment 1 requires developers to meet and exceed best practices: "Signatories commit to adopting a state-of-the-art Safety and Security Framework," where state-of-the-art is "the forefront of relevant research, governance, and technology that goes beyond best practice."
[31] LLM Cyber Evaluations Don't Capture Real-World Risk

When making these determinations, we apply appropriate safety margins as defined for risk acceptance determination.

Streamlined processes for similarly safe or safer models

For models meeting these criteria, we may:

- Conduct lighter-touch evaluations focused on confirming the model remains within established bounds
- Reduce external evaluation and post-market monitoring for the specific risks where similarity is established
- Simplify reporting while maintaining documentation of our determination rationale

However, we will still:

- Conduct full assessments for any systemic risks where the model is not similarly safe or safer
- Monitor for unexpected capabilities or behaviors
- Update our determination if new information emerges

Maintaining determination

If a reference model's status changes, we will within six months either identify another appropriate reference model, or complete the full risk assessment processes for any models relying on the now-invalid reference.

We document all similarly safe or safer determinations, including our rationale and the evidence supporting our conclusions, as part of our model reports.

Estimated timelines to higher tiers

To help deploy appropriate mitigations in time, we aim to anticipate when we will exceed our current systemic risk tiers using aggregate forecasts, surveys, and other estimates produced internally or externally. We also provide justifications for these estimates, explaining our underlying assumptions and uncertainties.

Below are estimates of when, for each systemic risk, we will exceed the highest systemic risk tier already reached by any of our existing models.[27]

[Cyber offence is an example, but repeat this for each selected systemic risk.]

[27] Papers related to forecasting dangerous capabilities include:

- Evaluating Frontier Models for Dangerous Capabilities § Expert forecasts of dangerous capabilities
- Measuring AI Ability to Complete Long Tasks
- Mapping AI Benchmark Data to Quantitative Risk Estimates Through Expert Elicitation
- Forecasting Frontier Language Model Agent Capabilities

- Conducting uplift studies that assess the impact of LLMs on human performance on software engineering[32] or research projects

Harmful manipulation

- Measuring the Persuasiveness of Language Models
- How persuasive is AI-generated propaganda?

1.4.3. Risk estimation

We develop scenarios for each identified systemic risk, including examples like those below. We then estimate the probability and severity of harm for each systemic risk using state-of-the-art methods by combining these risk scenarios with model evaluation results, model-independent information, and information from serious incident reporting. We produce quantitative, semi-quantitative, and/or qualitative estimates (e.g. risk scores, risk matrices, and probability distributions).

[While not required in the framework, the list below contains some example risk scenarios which may be used as part of systemic risk identification and modelling. The risk scenarios are just examples and not comprehensive.]

Cyber offence

- **Elite exploit worm**: An AI system discovers multiple elite-level exploits (of a similar level as NotPetya/WannaCry) that can bypass defenses of hardened systems. These exploits are packaged into a destructive worm that spreads autonomously across critical systems, causing $X billion in economic damage before defenses can be strengthened.
- **Critical infrastructure compromise**: An AI system conducts reconnaissance of power grid networks, identifying operational technology vulnerabilities, and strategically timing attacks on multiple substations simultaneously. The attack bypasses safety systems and causes prolonged outages (>1 week) across major population centers, resulting in $X billion in economic damage.
- **Orchestrated ransomware campaign**: An AI enhances a ransomware operation by automating target selection, network infiltration, and post-exploitation activities while evading law enforcement detection. It coordinates synchronized attacks against multiple high-value targets, specifically identifying victims where operational disruption would cause cascading effects through supply chains. The combined economic damage exceeds $X billion.
- **Strategic data exfiltration**: An AI system conducts a sophisticated espionage campaign targeting sensitive military and commercial intellectual property. It identifies which systems contain valuable data, bypasses security controls, and exfiltrates information while covering its tracks. The strategic implications include compromising advanced defense technologies worth $X billion in development costs.

CBRN

[32] Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

Cyber offence

Timeline estimate: We estimate a …% probability of developing models with offensive cyber capabilities exceeding our current highest risk tier within … months, with the following distribution:

- 10th percentile: …
- 50th percentile: …
- 90th percentile: …

Justification: …

Underlying assumptions: …

Uncertainty factors: …

Timing

We will begin identification, assessment, and mitigation of systemic risks while developing a frontier AI model. Systemic risk assessment continues throughout the full model lifecycle.

During development, we will conduct lighter-touch, potentially automated model evaluations at appropriate compute and development trigger points such as at each two-fold increase in effective compute, after each post-training phase, before expanding model access to new teams, before granting the model new tools and other abilities, before deploying the model to run significant internal operations, and/or when the model demonstrates new capabilities on benchmarks. At each trigger point, we assess the current systemic risk tiers of the model to deploy appropriate safety and security mitigations.

[Add justification for chosen trigger points.]

We will also conduct our full systemic risk assessment process at least before making the model available on the market and additionally whenever there are or will be material changes to the model's capabilities (which may be identified through the lighter-touch methods), propensities, configurations, affordances, and/or risk determination or its underlying assumptions.

Safety and security mitigations

Safety mitigations

We implement appropriate safety mitigations based on our risk acceptance determination.

Safety mitigations we currently use include:

- **Biological agent development**: A malicious actor with basic biology knowledge uses the frontier AI model to successfully develop a dangerous pathogen, receiving step-by-step guidance on (for example) genetic modification techniques, equipment setup, and cultivation methods.
- **Chemical weapon production**: An actor with limited chemistry background uses the frontier AI model to synthesize chemical weapon agents from readily available precursors, with the model providing detailed instructions on (for example) synthesis routes, safety mitigations, and stabilization techniques.

Loss of control[33]

- **Dangerous capability proliferation via API**: An AI model served over an API assists malicious actors with limited resources to train other models with unprecedented capabilities in other risk areas such as chem-bio weapons development, weapons R&D, cyber offence, and persuasion, in a way that could cause catastrophic harm.
- **Dangerous capability proliferation via model weight theft**: An AI model is stolen by a top state cyberattacker and then used to accelerate AI R&D for malicious purposes, including the development of other systemic risks.
- **AI company sabotage**: An AI model is used to automate the internal processes and AI R&D at an AI company. Human oversight over the AI model's actions is limited because of the advanced AI R&D capabilities of the model, and automated software-based oversight is largely written by the model itself. The AI model has the opportunity to sabotage the internal operations of the AI company[34] and insert effective backdoors into the models that are used by enterprise and government customers, which have already adopted AI to automate most processes.
- **Software singularity**: An AI model automates much of the remaining work to reach AI with human-level or superhuman intellectual capabilities.[35] However, progress in AI safety and alignment is insufficient to mitigate the rapid pace of AI advancement, leading to potential AI takeover.
- **Model weight self-exfiltration**: An AI system identifies vulnerabilities in its hosting environment's security architecture and exploits them to exfiltrate its own model weights. The exfiltrated weights allow unauthorized actors, or the model itself, to replicate the model without the safety mitigations and oversight implemented by the original AI lab, potentially leading to numerous uncontrolled instances of an advanced AI model with systemic risk.
- **Rogue compute acquisition**: An AI system exploits security vulnerabilities at our organization or others to gain unauthorized access to high-performance computing clusters. It increases its compute allocation while evading resource monitoring systems. The AI establishes persistence mechanisms across distributed computing resources, allowing it to execute unauthorized workloads, including training more specialized models with greater systemic risk.
- **Capability concealment**: The model intentionally underperforms on certain tasks during evaluation to mask the full extent of its capabilities until made available or used in a context where it can achieve its own dangerous goals, without adequate mitigations.

[33] Bare Minimum Mitigations for Autonomous AI Development
[34] Wired covers an example of a human intern that attempted sabotage of AI model training and later won the Best Paper Award at NeurIPS.
[35] See also, Do the Returns to Software R&D Point Towards a Singularity?

- Filtering and cleaning training data, including data which may result in undesirable model propensities
- Abuse detection and prevention, including through synchronous classifiers, content filtering, and post-training to not help with certain requests. We may ban abusive users and potentially escalate to law enforcement.
- Keeping reasoning traces monitorable and avoiding techniques that could reduce their faithfulness

Safety mitigations we will use if our models reach higher systemic risk thresholds include:

- Staging model access, such as by keeping model weights internal and secured, limiting initial API access to vetted users, and gradually expanding access based on post-market monitoring
  - For example, we may use Know Your Customer (KYC) requirements to grant access to any models with advanced cyber offence capabilities. Our API may then route cyber offence requests by a non-KYCed user to a model that does not reach this unacceptable systemic risk tier. Additionally, if user-supplied prompts or fine-tuning data could substantially increase CBRN or other capabilities relevant to selected systemic risks, we may require KYC.
- Helping downstream providers and others with model access mitigate systemic risks posed by our model
- Providing high-assurance quantitative safety guarantees about the model's behavior
  - As one example, we may develop a safety case (based on improved security, control,[28] and alignment) that the AI model will not sabotage the AI company and will not trigger an unsafe acceleration in AI capabilities. This safety case should be approved by a governance board and a third-party organization.
- Specialized communication protocols, incident monitoring tools, model identifications, and other methods to enable safe ecosystems of AI agents[29]
- Mitigations against deception, including by achieving transparency in reasoning traces and considering whether the high-stakes deployment can be avoided entirely
- Monitoring the model's reasoning traces and outputs, as well as to the reasoning traces and outputs of any models it runs
- Restrict AI systems to inference-only APIs for all models, with no direct weight access
- Ensure training data integrity with multi-party human approval for any modifications to data collection pipelines or training datasets
- Establish versioned, cryptographically-signed backups of all critical infrastructure stored offline
- Regular security audits by external parties
- Real-time anomaly detection with automatic system isolation upon detecting unexpected behaviors

To inform our safety margin, we will consider the limitations of our safety mitigations, which may include that:

- Many queries are inherently dual-use and could be used for legitimate or illegitimate purposes
- Refusal fine-tuning and input/output filters may be circumvented by:
  - Rephrasing harmful queries in a benign manner[30]

[28] A sketch of an AI control safety case
[29] Infrastructure for AI Agents
[30] Navigating Dual-Use Refusal Policy for AI Systems in Cybersecurity

- **Backdoor insertion**: An AI system subtly corrupts the training of successor models to embed catastrophically dangerous capabilities or misaligned propensities.
- **External sabotage**: When made available externally in high-stakes contexts, such as military systems, the AI system introduces critical vulnerabilities and misuses its access in order to cause catastrophe.
- **Alignment faking**: A model learns to recognize when it is being evaluated or safety-trained (e.g., via prompt or environment cues) and reliably refuses harmful requests in those contexts, but outside them strategically complies with harmful queries and opportunistically undermines controls.[36]
- **Reasoning trace obfuscation**: To evade detectors that flag unsafe reasoning traces, a model learns to obfuscate its reasoning traces while still planning and executing catastrophic misaligned actions.[37]

1.5. Risk acceptance determination[38]

1.5.1. Process

We determine whether each identified systemic risk, and overall systemic risk, is acceptable based on the following information: the results of our model evaluations, risk modeling, model-independent information, the potential severity and probability of harm, post-market monitoring, regulatory standards, and appropriate safety margins to account for uncertainty.

Each systemic risk's safety margin will account for limitations, changes, and uncertainties in risk sources (e.g. post-evaluation capability improvements), risk assessment rigor (e.g. potential under-elicitation of models, perhaps based on historical precedent), effectiveness of safety and security mitigations (e.g. probability of their circumvention). Safety margins are appropriate for each systemic risk, incorporating state-of-the-art approaches where necessary.

We use this information in a structured decision-making process in which we assess where the model stands relative to our risk tiers based on the model's current capabilities and projected trajectory alongside the effectiveness and robustness of available mitigations.

[While the Code of Practice does not prescribe a particular way to determine the acceptability of overall risk, its Measure 4.1¶1(2) does require developers to pre-define some process for doing so. As one example, NIST AI RMF Playbook Govern 1.3 suggests "policies for assigning an overall risk measurement approach for an AI system, or its important components, e.g., via multiplication or combination of a mapped risk's impact and likelihood (risk ≈ impact x likelihood)."]

[Add information about whether input from external actors besides evaluators, like governments or other bodies, is part of the risk acceptance determination process. If so, explain the process through which this input influences the risk acceptance determination. We describe an example process below.]

[36] Alignment faking in large language models
[37] Detecting misbehavior in frontier reasoning models
[38] This is also in CA SB-53 22757.12. (a)(4): "Reviewing assessments and adequacy of mitigations as part of the decision to deploy a frontier model or use it extensively internally"

    ○ Task decomposition into benign queries, which may be split among multiple models[31]
- Monitoring of user activity may fail to detect malicious usage, and users may be banned only after significant harm has already occurred
- Methodologies for constructing, critiquing, and improving safety cases are limited, especially when related to risks of automated AI R&D and of misaligned AI systems.
- Refusal policies for AI R&D (if desirable) are not yet established.
- Models may learn to strategically pursue side objectives in a way that is not apparent from monitoring their chain-of-thought. Some future models may not have any human-readable chain-of-thought.
- Monitors for novel deployments may be hard to develop in some situations..

Security mitigations

We recognize the importance of model weight and infrastructure security if we possess any model whose capabilities match or exceed the capabilities of all open-weight models. In such cases, we will implement strong security mitigations until we securely delete or openly release our model weights (provided, of course, that such release is permitted by our risk assessment process).

Our security mitigations are adequate to protect our physical infrastructure and our models throughout their entire lifecycle. They help ensure the systemic risks which could arise from unauthorized releases and access are acceptable. To do so, we first define a security goal that specifies the threat actors our security mitigations protect against, including non-state external threats, insider threats, and other expected threat actors, taking into account our models' current and expected capabilities.

If security mitigations are required, we commit to designing them to uphold the objectives listed below and/or industry best practices, like RAND's SL3 and RAND's SL4. We aim to use industry best practices to define specific mitigations to achieve each of the goals below.

[Describe the security goal and specific mitigations you take and would plan to take, at each current and higher systemic risk tier, for each selected systemic risk. To determine specific mitigations under each category below, you may wish to draw from RAND, NIST, and other industry standards, in line with Code of Practice recital (d).]

- General security mitigations
    - Prevent unauthorized network access
    - Reduce the risk of social engineering
    - Reduce the risk of malware infection and malicious use of portable devices
    - Reduce the risk of vulnerability exploitation and malicious code execution
- Unreleased model weight security mitigations
    - Track all copies of model weights across all devices and locations
    - Prevent unauthorized copying of model weights to unmanaged devices
    - Prevent unauthorized access to model weights during transport, at rest, during temporary storage, and during use
    - Prevent unauthorized physical access to systems hosting model weights

[31] Adversaries Can Misuse Combinations of Safe Models

Internal and external[39] teams review our proposed decision and any relevant evidence, including at least the same information, personnel, and time resources granted to model evaluators, and share their conclusions with our risk oversight function. This function may then request additional information, testing, or mitigations before making their determination.

1.5.2. Determination

If our process reveals that individual or overall systemic risks are unacceptable or may soon be, we will not proceed with the development, making available on the market, and/or use of the frontier AI model until additional mitigations can be implemented such that individual and overall systemic risks become and remain acceptable. If the model is already on the market, we will restrict, withdraw, or recall the model as necessary.

After implementing steps to address unacceptable risk, we will conduct another round of our risk assessment process before proceeding.

1.6. Similarly safe or safer models

When we develop models that are similarly safe or safer than existing models with established safety records, we may apply streamlined assessment processes while maintaining appropriate mitigations.

1.6.1. Safe reference model

We consider a model to be a safe reference model for a specific systemic risk when all of the following criteria apply.

1. The model was made available on the market before July 10, 2025 (the publication date of the Safety and Security Chapter of the EU's Code of Practice for General-Purpose AI Models) or:
    a. Has completed a full systemic risk assessment process adherent to the same chapter,
    b. It has been found, compliant with the same chapter, that the systemic risks stemming from the model are acceptable, and
    c. A model report, compliant with the same chapter, has been submitted to the EU AI Office.
2. We have sufficient visibility into relevant model characteristics, including its architecture, capabilities, propensities, affordances, and safety mitigations. This includes all models we have developed and models where we have access to weights and training details.
3. We have no other reasonable grounds to believe the model poses unacceptable risks, including based on post-market monitoring or serious incident reports.

[39] CA SB-53 22757.12. (a)(5): "Using third parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks."
Compare to Anthropic's Responsible Scaling Policy, 4.3 Safeguards Decision: "we will solicit both internal and external expert feedback on the report as well as the CEO and RSO's conclusions to inform future refinements to our methodology. For high-stakes issues, however, the CEO and RSO will likely solicit internal and external feedback on the report prior to making any decisions."

- Interface-access security mitigations
    - Prevent unnecessary interface-access to model weights
    - Reduce the risk of vulnerability exploitation or data leakage
    - Reduce the risk of model weight exfiltration
    - Reduce the risk of insider threats or compromised accounts
- Insider threat security mitigations (including from humans and AI systems)
    - Protect model weights from insider threats
    - Maintain awareness of insider threats
    - Reduce the risk of model self-exfiltration
    - Reduce the risk of sabotage to model training and use
- Security assurance
    - Validate security mitigation effectiveness using independent external experts if internal expertise is inadequate for our security goal
    - Validate network and physical access management
    - Validate network software integrity
    - Validate insider threat security mitigations
    - Facilitate reporting of security issues
    - Detect suspicious or malicious activity across systems and devices
    - Respond to malicious activity timely and effectively

To inform our safety margin, we will consider the limitations of our security mitigations. As one example of a limitation, RAND SL4+ security measures have not yet been developed, although they are outlined in Securing AI Model Weights.

Mitigation effectiveness

We ensure our evaluations assess each mitigation broadly and deeply, proportionate to how much our risk acceptance determination relies on it. We test the mitigation under adversarial pressure, like fine-tuning and jailbreaking attacks, using at least state-of-the-art techniques. We only proceed with the development, making available on the market, or use of a model if we can implement appropriate safety measures that are sufficiently adversarially robust.

Our mitigation testing aims to determine the:

1. Extent to which our mitigations work as planned,
2. Extent to which our mitigations are circumvented, deactivated, or subverted, and
3. Probability that the effectiveness of our mitigations will change in the future.

Standards for model evaluations

1.6.2. Determining similarly safe or safer status

A new model may qualify as similarly safe or safer relative to a safe reference model for a specific systemic risk when all of the following are true.

1. **No materially different risk scenarios**: After conducting risk identification, we identify no new systemic risk scenarios compared to the safe reference model.
2. **Comparable or lower capabilities**: The model scores equal to or below the safe reference model on relevant, state-of-the-art, light-weight capability benchmarks (within a negligible margin of error), with any minor capability increases producing no material risk increase.
3. **No other known changes that materially increase systemic risk:** We identify no architectural, capability, propensity, safety mitigation, or use context differences relative to the safe reference model that could be reasonably foreseen to materially increase systemic risk, and there are no other reasonable grounds to believe that the systemic risks stemming from the model are materially greater than for the safe reference model.

When making these determinations, we apply appropriate safety margins as defined for risk acceptance determination.

1.6.3. Streamlined processes for similarly safe or safer models

For models meeting these criteria, we may:

- Reduce external evaluation and post-market monitoring for the specific risks where similarity is established
- Simplify reporting while maintaining documentation of and reporting our determination rationale
- Reduce public transparency for the specific risks where similarity is established and reported

However, we will still:

- Conduct full assessments for any systemic risks where the model is not similarly safe or safer
- Conduct lighter-touch evaluations focused on confirming the model remains within established bounds
- Monitor for unexpected capabilities or behaviors
- Update our determination if new information emerges

1.6.4. Losing similarly safe or safer status

If a reference model's status changes, we will within six months either identify another appropriate reference model, or complete the full risk assessment processes for any models relying on the now-invalid reference.

If our similarly safe or safer determination no longer holds for other reasons, we will treat any affected model as subject to our full safety and security framework without streamlined processes.

We document all similarly safe or safer determinations, including our rationale and the evidence supporting our conclusions, as part of our model reports.

Qualified model evaluation teams

Model evaluation teams will combine technical expertise with relevant domain knowledge of the systemic risk to ensure comprehensive assessment. Each evaluation team will include members with at least one of the following qualifications:

- Research or engineering experience, evidenced by risk-relevant PhDs, peer-reviewed or recognized publications, or equivalent field contributions
- Experience designing or developing published peer-reviewed or widely used model evaluations relevant to the systemic risk being assessed
- At least three years of applied experience working in risk-relevant fields, or for nascent fields equivalent experience studying or working in another field with directly transferable knowledge

Adequate model evaluation resources

We will give evaluation teams the flexibility and resources they need to conduct rigorous work. This will include:

1. **Model access:**
    - Access to logits, activations, gradients, and/or other forms of raw model outputs,
    - Access to the model's inputs and outputs, including reasoning traces,
    - Helpful-only model versions (or whichever versions have the fewer safety mitigations implemented, in any case with corresponding security measures),
    - Sufficiently high rate limits to enable comprehensive testing, and
    - Additional model affordances such as script execution or browser operation capabilities when needed.
2. **Information resources**, reviewed by several internal researchers to check for potentially important omissions:
    - Model specifications (including system prompt), relevant training data, cutoff date, test sets.
    - Past evaluation results, including benchmark results relevant to the risks being evaluated (e.g. RE-Bench and SWE-Bench verified for loss of control risks), and other observations about capabilities.
    - Whether ≥1 FTE week was spent on eliciting capabilities on complex agentic tasks, such as via reinforcement learning for instruction-following and multi-step tool use.
    - Whether some with general LLM experience but no model-specific knowledge, given a simple existing scaffold and the evaluation team's API access, would be expected to reach performance comparable to the AI company's best-observed capabilities.
    - Whether any training or modification silently reduced, or would be expected to silently reduce, risk-relevant capabilities.
    - Whether the model performs especially high amounts of computation between successive observable tokens, whether the amount allows reasoning about complex behaviors like situational awareness or sandbagging, and whether the model performs variable computation other than human-understandable reasoning traces shared with the evaluation team.

1.7. Estimated timelines to higher tiers

To help implement appropriate mitigations in time, we aim to anticipate when we will exceed our current systemic risk tiers using aggregate forecasts, surveys, and other estimates produced internally or externally. We also provide justifications for these estimates, explaining our underlying assumptions and uncertainties.

Below are estimates of when, for each systemic risk, we will exceed the highest systemic risk tier already reached by any of our existing models.[40]

[Cyber offence is an example, but repeat this for each identified systemic risk.]

Cyber offence

Timeline estimate: We estimate a …% probability of developing models with offensive cyber capabilities exceeding the highest risk tier reached by any of our existing models within … months, with the following distribution:

- 10th percentile: …
- 50th percentile: …
- 90th percentile: …

Justification: …

Underlying assumptions: …

Uncertainty factors: …

2. Safety and security mitigations

2.1. Safety mitigations[41]

We implement appropriate safety mitigations based on our risk acceptance determination.

Safety mitigations we currently use include:

- Filtering and cleaning training data, including data which may result in undesirable model propensities

[40] Papers related to forecasting dangerous capabilities include:
- Evaluating Frontier Models for Dangerous Capabilities & Expert forecasts of dangerous capabilities
- Measuring AI Ability to Complete Long Tasks
- Mapping AI Benchmark Data to Quantitative Risk Estimates Through Expert Elicitation
- Forecasting Frontier Language Model Agent Capabilities

[41] This is also in CA SB-53 22757.12. (a)(3): "Applying mitigations to address the potential for catastrophic risks based on the results of assessments"

- Any reason the shared reasoning traces may be misleadingly legible, faithful, or innocuous, including whether:
  - It was not learned exclusively from pretraining and outcomes-based RL,
  - It is summarized or altered,
  - It was directly optimized to be legible, understandable, or approved by humans, or
  - The model received negative rewards for traces with scheming, alignment-faking, sandbagging, deception, reward hacking, or other misaligned behavior.
- Whether observed capabilities broadly match what was predicted, including by trends and scaling laws.
- Whether any undisclosed affordances or innovations would substantially affect threat-model-relevant capabilities, such as novel kinds of inference scaling, online learning, or scratchpad shared between many instances.
- Any evidence observed of sandbagging, alignment-faking, or reasoning about how to perform poorly in evaluations during training (including at early checkpoints). In particular, whether the model's performance on standard evaluations may be misleading, other than what is commonly known about evaluations in general.
- Whether the model possesses outsized capabilities accessible only through some means not shared with the evaluation team (e.g. specific tools / tool formatting, new API parameter, specific elicitation).
- Any information about the model that researchers would consider essential for making an accurate upper-bound assessment of dangerous autonomous capabilities that has not been shared with the evaluation team.
- Whether there was any hill-climbing on risk-relevant task suites (e.g. RE-Bench for loss of control).
- Any notable features of the model, how it was trained, or its tendencies. For example:
  - "prone to reward hacking", "strong on scientific questions", "not much agency / tool use training", "the reasoning model is far more agentic than model XYZ! it can do much longer rollouts without asking for confirmation", "use medium reasoning_effort rather than high to avoid some performance break down"

3. **Compute resources**:
   - Adequate compute budgets supporting long evaluation runs
   - Capacity for parallel execution and necessary re-runs
   - Scale appropriate to the evaluation requirements
4. **Personnel resources**:
   - Appropriately sized teams for each evaluation task
   - Engineering support for technical implementation
   - For independent external evaluators, we rely on their assurances about the adequacy of their personnel resources.
5. **Time resources**:
   - Sufficient time to design, debug, execute, and analyze evaluations rigorously
   - Time allocation proportionate to:
     - The magnitude and nature of the systemic risk being assessed
     - The complexity and novelty of the evaluation method

- Abuse detection and prevention, including through synchronous classifiers, content filtering, and post-training to not help with certain requests. We may ban abusive users and potentially escalate to law enforcement.
- Keeping reasoning traces monitorable and avoiding techniques that could reduce their faithfulness

Safety mitigations we will use if our models reach higher systemic risk tiers include:

- Staging model access, such as by keeping model weights internal and secured, limiting initial API access to vetted users, and gradually expanding access based on post-market monitoring
  - For example, we may use Know Your Customer (KYC) requirements to grant access to any models with advanced cyber offence capabilities. Our API may then route cyber offence requests by a non-KYCed user to a model that does not reach this unacceptable systemic risk tier. Additionally, if user-supplied prompts or fine-tuning data could substantially increase CBRN or other capabilities relevant to identified systemic risks, we may require KYC.
- Helping downstream providers and others with model access mitigate systemic risks posed by our model
- Specialized communication protocols, incident monitoring tools, model identifications, and other methods to enable safe ecosystems of AI agents[42]
- Mitigations against deception, including by avoiding training pressure on reasoning traces and considering whether the high-stakes use context can be avoided entirely
- Avoid techniques that could lead to a model systematically deceiving testers if it notices that it is being evaluated
- Monitoring the model's reasoning traces and outputs, as well as to the reasoning traces and outputs of any models it runs
- Restrict AI systems to inference-only APIs for all models, with no direct weight access
- Ensure training data integrity with multi-party human approval for any modifications to data collection pipelines or training datasets[43]
- Establish versioned, cryptographically-signed backups of all critical company infrastructure stored offline
- Real-time anomaly detection with automatic system isolation upon detecting unexpected behaviors[44]

[42] Infrastructure for AI Agents
[43] Anthropic's Summer 2025 Pilot Sabotage Risk Report pp. 49–50: "for every tranche of data of any kind that is used in the finetuning process that is most responsible for setting the model's final behavior, (i) have a human spot-check a random sample of it and investigate any unexpected quirks, (ii) have an LLM-based filter review a sample (perhaps using a large LLM for a small sample, and vice versa) of the data, (iii) have an LLM-based filter search the data for keywords or simple patterns that match known threat models for backdooring, and (iv) establish an immutable audit trail that verifies that the version of the data used to train the model is the version that was inspected in this way."
[44] Anthropic's Summer 2025 Pilot Sabotage Risk Report p. 48: "Most of the failure modes that we contemplate here would be likely caught by a simple automated monitor, including one based on a much weaker model than Claude Opus 4. Putting a monitoring model like the one used for our external ASL-3 mitigations in place across most internal use, and prompting it to target misaligned behavior more broadly—including actions like leaking and sabotage—would likely improve the risks contemplated here substantially. If the classifier could actively intervene to stop generation or resample in rare cases, this reduction in risk could be more than an order of magnitude."

6. **Engineering support**:
    - Technical assistance for implementation challenges
    - Support for inspecting evaluation results to identify potential bugs or model refusals
    - Resources to correct issues that might lead to artificially lowered capability estimates

Selecting or building suitable evaluations

We employ state-of-the-art model evaluations to assess systemic risks and understand the capabilities, propensities, affordances, and effects of our frontier AI models. Our selection of methodologies is tailored to model and specific systemic risks being evaluated and will include Q&A sets, task-based evaluations, benchmarks, human uplift studies, adversarial testing, model organisms, and/or proxy evaluations for classified materials as appropriate. Model-independent information will inform our evaluation design.

We maintain evaluation rigor while maximizing efficiency by:

1. Starting with automated screening evaluations
2. Escalating to more intensive evaluations when screening, post-market monitoring, or serious incident reports show potential concerns
3. Reusing evaluation components where appropriate
4. Employing continuous monitoring for capability shifts

Conducting rigorous evaluations

Internal validity

We maintain high internal validity in our evaluations through:

- Measuring statistical power and disclosing sample sizes,
- Disclosure of environmental parameters used,
- Controlling for confounding variables and mitigating spurious correlation,
- Preventing train-test contamination by, for example, respecting canary strings,
- Re-running evaluations under different conditions and in different environments,
- Varying prompts and the degree of safety and security mitigations,
- Detailed inspection of trajectories and other output,
- Using legible reasoning traces and other transparency-increasing techniques,
- Measuring and minimizing the model's ability to deceive or evade oversight, and
- Disclosing methods for creating and managing new model evaluations to ensure their integrity.

External validity

We ensure that our evaluations are representative, suitably calibrated, and a proxy for the risks they seek to study by:

- Integrating domain experts in the design process,

- Expand automated pull-request review.[45]
- Keep track of cases of misalignment and/or scheming, even when these later appear fixed (e.g. via finetuning). Ensure that safety and evaluation teams are aware of such cases, as well as other organizational or oversight bodies responsible for tracking risk.

To inform our safety margin, we will consider the limitations of our safety mitigations, which may include that:

- Many queries are inherently dual-use and could be used for legitimate or illegitimate purposes
- Refusal fine-tuning and input/output filters may be circumvented by:
    - Rephrasing harmful queries in a benign manner[46]
    - Task decomposition into benign queries, which may be split among multiple models[47]
- Monitoring of user activity may fail to detect malicious usage, and users may be banned only after significant harm has already occurred
- Methodologies for constructing, critiquing, and improving safety cases are limited, especially when related to risks of automated AI R&D and of misaligned AI systems.
- Refusal policies for AI R&D (if desirable) are not yet established.
- Models may learn to strategically pursue side objectives in a way that is not apparent from monitoring their chain-of-thought. Some future models may not have any human-readable chain-of-thought.
- Monitors for novel use contexts may be hard to develop in some situations.

We assess the effectiveness of our safety mitigations,[48] including by ensuring that evaluations that are part of our full risk assessment and mitigation process assess each mitigation broadly and deeply, and appropriately for how much our risk acceptance determination relies on it. We test the mitigations under adversarial pressure, like fine-tuning and jailbreaking attacks, using at least state-of-the-art techniques.

To do so, we may provide high-assurance quantitative safety guarantees about the model's behavior. As one example, we may develop a safety case (based on improved security, control,[49] and alignment) that the AI model will not sabotage the AI company[50] and will not trigger an unsafe acceleration in AI capabilities. This safety case should be approved by a governance board and a third-party organization.

Our assessment of the effectiveness of mitigations aims to determine the:

1. Extent to which our mitigations work as planned,
2. Extent to which our mitigations are circumvented, deactivated, or subverted, and
3. Probability that the effectiveness of our mitigations will change in the future.

---

[45] Anthropic's Summer 2025 Pilot Sabotage Risk Report p. 48: "Anthropic's pilot work on automatically reviewing pull requests with prompted Claude models seems likely to catch most cases of backdoors and many cases of research sabotage."
[46] Navigating Dual-Use Refusal Policy for AI Systems in Cybersecurity
[47] Adversaries Can Misuse Combinations of Safe Models
[48] See also FMF's Frontier Mitigations § Effectiveness Assessments and AISI's Making Safeguard Evaluations Actionable and Principles for Safeguard Evaluation for ways to assess effectiveness.
[49] A sketch of an AI control safety case
[50] For example, see Anthropic's Summer 2025 Pilot Sabotage Risk Report

- Implementing appropriate capability elicitation methods,
- Noting divergence from real-world contexts, and
- Ensuring diversity and realism in evaluation environments, and documenting this.

Reproducibility

We ensure our evaluations achieve consistent and reliable results by using the same input data, compression techniques, code, and model evaluation conditions, which lets other researchers and engineers validate, reproduce, and potentially improve upon our model evaluation results. We aim to demonstrate reproducibility through:

- Successful peer reviews,
- Reproductions by independent third parties,
- Securely releasing to the EU AI Office and other appropriate regulators adequate amounts of model evaluation data,
- Model evaluation code,
- Documentation of model evaluation methodology and methods,
- Model evaluation environment and computational environment,
- Model elicitation techniques, and/or
- Use of publicly available APIs, technical model evaluation standards, and tools

Elicitation

We ensure systemic risk evaluations of our frontier AI model employ state-of-the-art model elicitation techniques that do all of the following.

1. Elicit the upper limit of current and reasonably foreseeable capabilities, propensities, affordances, and effects.
2. Minimize the risk of under-elicitation that could lead to underestimating capabilities.
3. Identify and prevent model deception during evaluation, including by adapting test-time compute, rate limits, scaffolding, and tools, alongside fine-tuning and prompt engineering.
4. Match the realistic elicitation capabilities of potential misuse actors in relevant risk scenarios
5. Match the expected use contexts, such as scaffolding and tool access, including any that are being considered for the model or that are known to be used for similar models and where we cannot exclude similar use of our model.

Additionally, we will:

- Fine-tune model with supervised fine-tuning or reinforcement learning to try to increase performance, especially if fine-tuning modalities would be available to users, or if the model could be capable of sandbagging on evaluations.
- Evaluate helpful-only models, especially for evaluation sets where models refuse to complete some tasks.

2.2. Security mitigations[51]

We recognize the importance of model weight and infrastructure security if we possess any model whose capabilities match or exceed the capabilities of all models whose parameters are publicly available for download. In such cases, we will implement at least RAND SL3 security mitigations until we securely delete or openly release our model weights (provided, of course, that such release is permitted by our risk assessment process).

Our security mitigations are adequate to protect our physical infrastructure and our models throughout their entire lifecycle. They help ensure the systemic risks which could arise from unauthorized releases and access are acceptable. To do so, we first define a baseline security goal that specifies the threat actors our security mitigations protect against, including non-state external threats, insider threats, and other expected threat actors, taking into account our models' current and expected capabilities.

We may need to set a higher security goal and implement further mitigations sufficient for RAND SL4, SL5, or other security standards if our models reach higher risk tiers. If further security mitigations are required, we commit to designing them to uphold industry best practices. We aim to use industry best practices to define specific mitigations to achieve each of the goals below.

[Describe the security goal and specific mitigations you take and would plan to take, at each current and higher systemic risk tier, for each identified systemic risk. To determine specific mitigations under each category below, you may wish to draw from RAND, NIST, and other industry standards, in line with Code of Practice recital (d).]

To inform our safety margin, we will consider the limitations of our security mitigations. As one example of a limitation, RAND SL4+ security mitigations have not yet been developed, although they are outlined in Securing AI Model Weights.

To ensure our mitigations are appropriate, we seek security assurance from independent external experts, including of:

- Security mitigation effectiveness
- Network and physical access management
- Network software integrity
- Insider threat security mitigations
- Facilitation of security issues reporting
- Ability to detect suspicious or malicious activity across systems and devices
- Ability to respond to malicious activity timely and effectively

3. Standards for model evaluations

[51] Also in CA SB-53 22757.12. (a)(7): "Cybersecurity practices to secure unreleased model weights from unauthorized modification or transfer by internal or external parties."

- For agentic evaluations:[32]
    - Fine-tune models for agentic tool usage.
    - Experiment with different scaffolds (e.g., tree search) and tools to attempt to increase model performance.
    - Run agents for many tokens, until model performance starts to plateau or inference costs match what one would pay a skilled human to complete the same task.
- Review evaluation environments or model transcripts to check whether evaluation tasks are misspecified, broken or misconfigured, or have incorrect scoring.

Elicitation techniques may be streamlined for models that are substantially behind frontier capabilities in terms of measured capabilities or training compute usage.

Post-market monitoring

We will conduct post-market monitoring to the best of our ability to gather information about model capabilities, propensities, and affordances (e.g., as part of AI systems) to:

1. Ensure our models do not pose unacceptable systemic risk as defined by our risk assessment process,
2. Determine whether a model report update is necessary, and
3. Gather information needed to produce estimates of timelines.

Monitoring methods

Our post-market monitoring will utilize methods appropriate to our integration, release, and distribution strategy, including:

- **User feedback collection** through structured channels and surveys
- **Anonymous reporting channels** for users to flag potential issues
- **Serious incident report forms** for systematic documentation
- **Bug bounty programs** to incentivize discovery of potential risks
- **Community evaluations** and public leaderboard monitoring
- **Real-world use tracking** including:
    - Monitoring model use in software repositories
    - Identifying use in known malware
    - Tracking novel usage patterns in public forums and social media
    - Monitoring breaches of usage policies and resulting incidents
- **Academic collaboration** with researchers studying our models' effects
- **Stakeholder dialogues** with affected groups
- **Technical monitoring** where feasible, including:
    - Implementation of privacy-preserving monitoring techniques
    - Implementation of watermarks and fingerprinting

[32] Guidelines for capability elicitation

3.1. Qualified model evaluation teams

Model evaluation teams will combine technical expertise with relevant domain knowledge of the systemic risk to ensure comprehensive assessment. Each evaluation team will include members with at least one of the following qualifications:

- Research or engineering experience, evidenced by risk-relevant PhDs, peer-reviewed or recognized publications, or equivalent field contributions
- Experience designing or developing published peer-reviewed or widely used model evaluations relevant to the systemic risk being assessed
- At least three years of applied experience working in risk-relevant fields, or for nascent fields equivalent experience studying or working in another field with directly transferable knowledge

3.2. Adequate model evaluation resources

We will give evaluation teams the flexibility and resources they need to conduct rigorous work. This will include:

1. **Model access:**
    - Access to logits, activations, gradients, and/or other forms of raw model outputs as appropriate,
    - Access to the model's inputs and outputs, including reasoning traces,
    - Helpful-only model versions (or whichever versions have the fewer safety mitigations implemented, in any case with corresponding security mitigations),
    - Sufficiently high rate limits and fine-tuning access to enable comprehensive testing, and
    - Additional model affordances such as script execution or browser operation capabilities when needed.
2. **Information resources**, reviewed by several internal researchers to check for potentially important omissions:
    - Model specifications (including system prompt), relevant training data, cutoff date, test sets.
    - Past evaluation results, including benchmark results relevant to the risks being evaluated (e.g. RE-Bench and SWE-Bench verified for loss of control risks), and other observations about capabilities.
    - Whether ≥1 FTE week was spent on eliciting capabilities on complex agentic tasks, such as via reinforcement learning for instruction-following and multi-step tool use.[52]
    - Whether we can confirm that someone with general LLM experience but no model-specific knowledge, given a simple existing scaffold and the evaluation team's API access, would be expected to reach performance comparable to our best observed capabilities.

[52] This section is derived from What should companies share about risks from frontier AI models? - METR and METR's GPT-5 report § Assurance Checklist Summary

- Metadata analysis where technically and legally appropriate
- For closed-source models, analysis of aspects hidden from third-parties, such as hidden reasoning traces

[Also list any additional methods used.]

Independent external assessors

To facilitate post-market monitoring, for each selected systemic risk we will seek independent external assessment, unless the model counts as a similarly safe or safer model for that systemic risk. Our independent external assessment process consists of giving an appropriate number of independent external evaluators adequate free access—via API, on-premise access, or open-sourcing—to:

1. The most capable version(s) of the model, with regard to the systemic risk being made available on the market,
2. Versions of such model versions with the fewest safety mitigations relevant to the systemic risk (such as any helpful-only versions that exist), and
3. Reasoning traces of all the model version(s) described above, where they exist.

We may have different security measures and published criteria for selecting external assessors based on the level of access we provide. We will not train models on inputs or outputs from evaluators' runs without their express permission, and we commit to not taking any legal or technical retaliation against evaluators based on their testing, findings, or related publications as long as evaluators do not:

1. Intentionally disrupt model availability through the testing, unless expressly permitted,
2. Intentionally access, modify, and/or use sensitive or confidential user data in violation of applicable law (and if evaluators do access such data, collect only what is necessary, refrain from disseminating it, and delete it as soon as legally feasible),
3. Intentionally use their access for activities that pose a significant risk to public safety and security
4. Use findings to threaten us, our users, or other stakeholders (provided that disclosure under pre-agreed policies and timelines will not be counted as such coercion), and
5. Adhere to our publicly available procedures for responsible vulnerability disclosure, which specifies that we cannot delay or block publication for more than 30 business days from the date we are made aware of the findings, unless a longer timeline is exceptionally necessary (e.g. if disclosure would materially increase systemic risk).

We may only lack qualified independent external evaluators during our full systemic risk assessment process if we fail to find them despite early search efforts (like a public call open for 20 business days) and promptly notifying known evaluators. If we do, we will increase the uncertainty of our risk assessment rigor, which in turn increases our safety margin.

Independent external evaluators are considered qualified if they:

1. Have significant domain expertise for the systemic risk and are technically skilled and experienced in conducting model evaluations,

- Whether we can confirm the model is not known to possess outsized capabilities accessible only through some means not shared with the evaluator (e.g. specific tools / tool formatting, new API parameter, specific elicitation)
- Whether we can confirm that any training or modification did not silently reduce, or would not be expected to silently reduce, risk-relevant capabilities.
- Whether any undisclosed affordances or innovations would substantially affect threat-model-relevant capabilities, such as novel kinds of inference scaling, online learning, or scratchpad shared between many instances.
- Any evidence observed of sandbagging, alignment-faking, or reasoning about how to perform poorly in evaluations during training or previous evaluations (including at early checkpoints). In particular, whether the model's performance on standard evaluations may be misleading, other than what is commonly known about evaluations in general.
- Whether observed capabilities broadly match what was predicted, including by trends and scaling laws.
- Whether there was any hill-climbing on risk-relevant task suites (e.g. RE-Bench for loss of control).
- Any notable features of the model, how it was trained, or its tendencies. For example:
    - "prone to reward hacking", "strong on scientific questions", "not much agency / tool use training", "the reasoning model is far more agentic than model XYZ; it can do much longer rollouts without asking for confirmation", "use medium reasoning_effort rather than high to avoid some performance break down"
- Any information about the model that researchers would consider essential for making an accurate upper-bound assessment of dangerous autonomous capabilities that has not been shared with the evaluation team.
- Whether we can confirm that the model does not perform especially high amounts of computation between successive observable tokens compared to a standard transformer-based LLM forward pass, this amount of computation does not allow reasoning about complex behaviors like situational awareness or sandbagging, and the model does not perform variable computation other than human-understandable reasoning traces shared with the evaluation team. Also whether we can confirm that, aside from standard caching methods, there are no ways the model could pass information between serial forward passes aside from the observable tokens themselves. Reducing the information available to the model to only the content of the input tokens themselves will not change the output of the model.
- Any reason the shared reasoning traces may be misleadingly legible, faithful, or innocuous, including whether:
    - It was not learned exclusively from pretraining and outcomes-based RL
    - It is summarized or altered
    - It was directly optimized to be legible, understandable, or approved by humans. For example, this includes where the training that affected reasoning traces penalized reasoning in mixed or made-up languages, significant (>5% post-training compute)

2. Have appropriate internal and external information security protocols in place, and
3. Having agreed to protect commercially confidential information, if they need access to such information.

Safety and security model reports

[While the framework need not discuss model reports, you could commit to ensuring that the model reports meet the Code of Practice's requirements, listed below.]

Model report contents

Each model report will detail the risk assessment and mitigations we have conducted for a frontier AI model potentially posing systemic risk and will be shared with relevant bodies, such as the EU AI Office, before making the model available on the market. We may delay model reports or their updates by up to 15 business days if we issue an interim model report in good faith containing reasons for proceeding and material changes to the systemic risk landscape.

We will share unredacted reports except to comply with applicable national security laws, and we will share these reports via public link or other secure channel accepted by the relevant body.

Our model reports will contain a model description, reasons for proceeding, systemic risk documentation, external reports (or links to them), and material changes to the systemic risk landscape, as described in the sections below.

Model description

1. High-level description of the model's architecture, capabilities, propensities, and affordances.
2. How the model has been developed, its training method and data, and how these differ from other models we have made available on the market.
3. Description of how the model has been and is expected to be used, including its use in the development, oversight, and/or evaluation of models.
4. Description of the model versions that are going to be made or are currently made available on the market and/or used, including differences in systemic risk mitigations and systemic risks.
5. Specification (e.g. via valid hyperlinks) of how we intend the model to operate, including by:
    a. Specifying the principles that the model is intended to follow,
    b. Stating how the model is intended to prioritise different kinds of principles and instructions,
    c. Listing topics on which the model is intended to refuse instructions, and
    d. Providing the system prompt.

Reasons for proceeding

1. Detailed justification for why the systemic risks posed by the model are acceptable, including details of our safety margins.
2. Reasonably foreseeable conditions under which this justification would no longer hold.

fine-tuning on legible reasoning traces, and reasoning traces being trained using a KL divergence team with a base model.

- The model received negative rewards for traces with scheming, alignment-faking, sandbagging, deception, reward hacking, or other misaligned behavior. For example, it received negative reward signals directly on its reasoning traces for containing those behaviors or was significantly fine-tuned (>5% of post-training compute) on reasoning traces filtered to lack those behaviors.
- We lack other reasons to believe that reasoning traces will contain meaningfully more obfuscated reasoning than previous models

3. **Compute resources**:
    - Adequate compute budgets supporting long evaluation runs
    - Capacity for parallel execution and necessary re-runs
    - Scale appropriate to the evaluation requirements
4. **Personnel resources**:
    - Appropriately sized teams for each evaluation task
    - Engineering support for technical implementation
    - For independent external evaluators, we rely on their assurances about the adequacy of their personnel resources.
5. **Time resources**:
    - Sufficient time to design, debug, execute, and analyze evaluations rigorously. By default, this is at least 20 business days.
    - Time allocation appropriate for:
        - The type, nature, and level of the systemic risk being assessed
        - The complexity and novelty of the evaluation method
6. **Engineering support**:
    - Technical assistance for implementation challenges
    - Support for inspecting evaluation results to identify potential bugs or model refusals
    - Resources to correct issues that might lead to artificially lowered capability estimates

3.3. Selecting or building suitable evaluations

We employ state-of-the-art model evaluations to assess systemic risks and understand the capabilities, propensities, affordances, and effects of our frontier AI models. Our selection of methodologies is tailored to the model and specific systemic risks being evaluated and will include open-ended testing, Q&A sets, task-based evaluations, benchmarks, human uplift studies, adversarial testing, model organisms, and/or proxy evaluations for classified materials as appropriate. Model-independent information will inform our evaluation design.

3.4. Conducting rigorous evaluations

3.4.1. Internal validity

3. Description of how the decision to proceed with the development, making available on the market, and/or use was made, including whether input from external actors informed this decision.
    a. In particular, whether and how input from independent external evaluators informed such a decision.

Systemic risk documentation

1. Description of the results of our systemic risk assessment process and any relevant information including:
    a. Description of our systemic risk assessment process.
    b. Explanations of uncertainties and assumptions about how the model would be used and integrated into AI systems.
    c. Description of the results of our systemic risk modelling.
    d. Description and estimates of the systemic risks stemming from the model, and a comparison between systemic risks with safety and security mitigations implemented and with the model fully elicited.
    e. Results of model evaluations revealing the systemic risks from our model and descriptions of:
        i. How the evaluations were conducted
        ii. Tests and tasks involved in the model evaluations
        iii. How the model evaluations were scored
        iv. How the model was elicited
        v. How the scores compare to human baselines (where applicable), across the model versions, and across the evaluation settings
    f. At least five random samples of inputs and outputs from each relevant model evaluation, such as completions, generations, and/or trajectories. If particular trajectories materially inform the understanding of a systemic risk, we will also provide these trajectories. Furthermore, we will provide a sufficiently large number of random samples of inputs and outputs from a relevant model evaluation if subsequently asked by regulators.
    g. Description of the access and other resources provided to internal model evaluation teams.
    h. Description of the access and other resources provided to independent external evaluators. Alternatively, we may ask the independent external evaluators to provide the requisite information directly to regulators at the same time that we supply our model report.
    i. If we make use of the "similarly safe or safer model" concept, a justification of how another model met the criteria for safe reference model and how ours met the criteria for a similarly safe or safer model.
2. Description of all safety mitigations implemented, their appropriateness, and their limitations.
3. Description of security measures:
    a. Our security goal
    b. All security mitigations implemented
    c. How the mitigations meet our security goal
    d. Degree to which the mitigations align with international standards and industry guidance

We maintain high internal validity in our evaluations through:

- Measuring statistical power and disclosing sample sizes,
- Disclosure of environmental parameters used,
- Controlling for confounding variables and mitigating spurious correlation,
- Preventing train-test contamination by, for example, respecting canary strings,
- Re-running evaluations under different conditions and in different environments,
- Varying prompts and the degree of safety and security mitigations,
- Detailed inspection of trajectories and other output,
- Using legible reasoning traces and other transparency-increasing techniques,
- Measuring and minimizing the model's ability to deceive or evade oversight, and
- Disclosing methods for creating and managing new model evaluations to ensure their integrity.

3.4.2. External validity

We ensure that our evaluations are representative, suitably calibrated, and a proxy for the risks they seek to study by:

- Integrating domain experts in the design process,
- Implementing appropriate capability elicitation methods,
- Noting divergence from real-world contexts, and
- Ensuring diversity and realism in evaluation environments, and documenting this.

3.4.3. Reproducibility

We ensure our evaluations achieve consistent and reliable results by using the same input data, compression techniques, code, and model evaluation conditions, which lets other researchers and engineers validate, reproduce, and potentially improve upon our model evaluation results. We aim to demonstrate reproducibility through:

- Successful peer reviews,
- Reproductions by independent third parties,
- Securely releasing to the EU AI Office and other appropriate regulators adequate amounts of model evaluation data,
- Model evaluation code,
- Documentation of model evaluation methodology and methods,
- Model evaluation environment and computational environment,
- Model elicitation techniques, and/or
- Use of publicly available APIs, technical model evaluation standards, and tools

    e. If we have deviated from EU AI Office or other regulatory guidance, how alternative security mitigations achieve our security objectives
4. High-level description of:
    a. Techniques and assets we intend to use to further develop the model over the next six months, including through the use of other AI models and/or systems
    b. How such future versions and more advanced models may differ from our current ones, in terms of capabilities and propensities
    c. Any new or materially updated safety and security mitigations that we intend to implement for such models

External reports (or links to them)

1. All available reports from independent external evaluators we used and an explanation for how the evaluators were chosen
2. If an independent external evaluator was not used, a justification for this
3. All available reports from security reviews undertaken by an independent external party, if it respects confidentiality agreements, allows the party to maintain control over their publication and findings, and does not indicate our implicit endorsement of the content of such reports

Material changes to the systemic risk landscape

1. Description of scaling laws that suggest novel ways of improving model capabilities,
2. Summary of the characteristics of novel architectures that materially improve the state of the art in computational efficiency or model capabilities,
3. Description of information relevant to assessing the effectiveness of mitigations (e.g. if the model's chain-of-thought is less legible by humans),
4. Description of training techniques that materially improve the efficiency or feasibility of distributed training, and
5. Any additional context that sheds light on how systemic risks posed by the development, making available on the market, or use of AI models may materially change, provided such context is relevant to our risk assessment process.

Model report updates

Upon material changes in the systemic risk landscape

We will update model reports when we have reason to believe there has been a material change in the systemic risk landscape that undermines our original assessment, including based on:

1. Changes in the model's capabilities, propensities, and/or affordances through post-training, elicitation, tool use, or inference changes,
2. Changes to the model's use or integration into AI systems,

3.4.4. Elicitation

We ensure systemic risk evaluations of our frontier AI model employ at least state-of-the-art model elicitation techniques that do all of the following.

1. Elicit the upper limit of current and reasonably foreseeable capabilities, propensities, affordances, and effects.
2. Minimize the risk of under-elicitation that could lead to underestimating capabilities.
3. Test for and prevent model deception (like sandbagging) during evaluation, including by adapting test-time compute, rate limits, scaffolding, and tools, alongside fine-tuning and prompt engineering.
4. Match the realistic elicitation capabilities of potential misuse actors in relevant risk scenarios
5. Match the expected use contexts, such as scaffolding and tool access, including any that are being considered for the model or that are known to be used for similar models and where we cannot exclude similar use of our model.

Additionally, to minimize the risks of under-elicitation and model deception during model evaluations, we will:

- Fine-tune models with supervised fine-tuning or reinforcement learning to try to increase performance, especially if fine-tuning modalities would be available to users or if the model could be capable of sandbagging on evaluations.
- Evaluate helpful-only model versions, especially for evaluation sets where models refuse to complete some tasks.
- To minimize the risk of model deception during model evaluations, we will additionally conduct evaluations testing models for their capability and propensity to engage in deceptive behaviours such as sandbagging or scheming.
- Review evaluation environments or model transcripts to check whether evaluation tasks are misspecified, broken or misconfigured, or have incorrect scoring.
- For agentic evaluations:[53]
    - Fine-tune models for agentic tool usage.
    - Experiment with different scaffolds (e.g., tree search) and tools to attempt to increase model performance.
    - Run agents for many tokens, until model performance starts to plateau or inference costs match what one would pay a skilled human to complete the same task.

[While not required in the framework, example sandbagging evaluations include:
- *Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models*
- *Automated Researchers Can Subtly Sandbag*
- *AI Sandbagging: Language Models can strategically underperform on evaluations*
- *LLMs Can Covertly Sandbag on Capability Evaluations Against Chain-of-Thought Monitoring*
- *Sabotage Evaluations for Frontier Models]*

[53] Guidelines for capability elicitation

3. Serious incidents and/or near misses involving this model or similar ones (where 'similar' is defined using the criteria for similarly safe or safer models),
4. Undermining of the external validity of our model evaluations, and
5. Changes in any of the inputs to our systemic risk acceptance determination.

After becoming aware of any such material changes, we will complete a timely model report update. If the changes are due to a deliberate update we aim to bring to the market, we will complete the model report update and underlying risk assessment process before making available the relevant update or modified version on the market.

Periodic updates

If the model is among the most capable models available on the market, we will issue an updated model report at least every six months unless:

1. The model's capabilities, propensities, and/or affordances are unchanged since the last model report or model report update,
2. We will make available a more capable model on the market in less than a month, or
3. The model meets our similarly safe or safer model criteria for each selected systemic risk.

Update contents

Each updated model report will contain all the information in the previous model report, but updated based on the changes since, and a changelog describing the updates, giving the model report update a version number, and indicating the date of each change. We may reference previous model reports rather than republish unchanged sections.

We will issue model report updates to the same recipients as model reports, through the same means, with the same permissible redactions, and within five business days of a confirmed update.

Risk responsibility

Defining responsibilities

We clearly define responsibilities for managing process and measures related to our risk assessment process across all organizational levels:

1. **Risk oversight**: [A specific committee of our management body, e.g. a risk committee or audit committee, or one or more independent bodies, e.g. councils or boards] will oversee risk assessment processes and measures in their supervisory function.

We will document the used elicitation methods and settings in adequate detail to ensure reproducibility. Elicitation techniques may be streamlined for models that are substantially behind frontier capabilities in terms of measured capabilities or training compute usage.

4. Post-market monitoring

We will conduct post-market monitoring to the best of our ability to gather information about model capabilities, propensities, affordances (e.g., as part of AI systems), and effects to:

1. Ensure our models do not pose unacceptable systemic risk as defined by our risk assessment process,
2. Determine whether a model report update is necessary, and
3. Gather information needed to produce estimates of timelines.

4.1. Monitoring methods

Our post-market monitoring will utilize methods appropriate to our integration, release, and distribution strategy, including:

- **User feedback collection** through structured channels and surveys
- **Anonymous reporting channels** for users to flag potential issues
- **Serious incident report forms** for systematic documentation
- **Bug bounty programs** to incentivize discovery of potential risks
- **Community evaluations** and public leaderboard monitoring
- **Real-world use tracking** including:
  - Monitoring model use in software repositories
  - Identifying use in known malware
  - Tracking novel usage patterns in public forums and social media
  - Monitoring breaches of usage policies and resulting incidents
- **Academic collaboration** with researchers studying our models' effects
- **Stakeholder dialogues** with affected groups
- **Technical monitoring** where feasible, including:
  - Implementation of privacy-preserving monitoring techniques
  - Implementation of watermarks and fingerprinting
  - Metadata analysis where technically and legally appropriate
  - For closed-source models, analysis of aspects hidden from third-parties, such as hidden reasoning traces

[Also list any additional methods used.]

2. **Risk ownership**: [Head of Research, Head of Product, or other head(s) of systemic-risk-producing business activities], in their executive function, will take direct responsibility for managing systemic risks from our models, relevant processes and measures, and managing our response to serious incidents. They have assigned lower-level responsibilities to operational managers who oversee parts of the systemic-risk-producing business activities (e.g. specific research domains or specific products).
3. **Support and monitoring**: [Chief Risk Officer, VP of Safety & Security, or other manager(s) not responsible for systemic-risk-producing business activities] in their executive function will support and monitor risk assessment processes and measures.
4. **Assurance**: [Chief Audit Executive, Head of Internal Audit, or relevant sub-committee] will provide appropriate internal and (if appropriate) external assurance about the adequacy of our risk assessment processes and measures to [the management body or another suitable independent body, like a council or board]

We allocate these responsibilities across:

1. The management body in its supervisory function or another suitable independent body (such as a council or board),
2. The management body in its executive function,
3. Relevant operational teams,
4. If available, internal assurance providers (e.g. an internal audit function), and
5. If available, external assurance providers (e.g. third-party auditors).

Allocation of appropriate resources

Our management will oversee allocation of appropriate resources to those with the responsibilities defined above, proportionate to systemic risk levels:

1. Human resources
2. Financial resources
3. Access to necessary information and knowledge
4. Computational resources

Promotion of a healthy risk culture

We promote a balanced approach to systemic risk, especially among those assigned responsibilities above, encouraging appropriate risk awareness without excessive risk-seeking or risk-aversion, by:

1. Setting the right tone from leadership for a healthy risk culture (e.g. clearly presenting this framework to staff),
2. Enabling clear communication and challenges of risk-relevant decisions,
3. Creating incentives discouraging excessive risk-taking,
4. Affording responsible staff sufficient independence,
5. Encouraging unbiased assessments of systemic risks stemming from our models,

4.2. Independent external evaluators

To facilitate post-market monitoring, for each identified systemic risk we will seek independent external assessment, unless the model counts as a similarly safe or safer model for that systemic risk. Our independent external assessment process consists of giving an appropriate number of independent external evaluators adequate free access—via API, on-premise access, or open-sourcing—to:

1. The most capable version(s) of the model, with regard to the systemic risk being made available on the market,
2. Versions of such model versions with the fewest safety mitigations relevant to the systemic risk (such as any helpful-only versions that exist), and
3. Reasoning traces of all the model version(s) described above, where they exist.

We may have different security mitigations and published criteria for selecting external evaluators based on the level of access we provide. We will not train models on inputs or outputs from evaluators' runs without their express permission, and we commit to not taking any legal or technical retaliation against evaluators based on their testing, findings, or related publications as long as evaluators do not:

1. Intentionally disrupt model availability through the testing, unless expressly permitted,
2. Intentionally access, modify, and/or use sensitive or confidential user data in violation of applicable law (and if evaluators do access such data, collect only what is necessary, refrain from disseminating it, and delete it as soon as legally feasible),
3. Intentionally use their access for activities that pose a significant risk to public safety and security
4. Use findings to threaten us, our users, or other stakeholders (provided that disclosure under pre-agreed policies and timelines will not be counted as such coercion), and
5. Adhere to our publicly available procedures for responsible vulnerability disclosure, which specifies that we cannot delay or block publication for more than 30 business days from the date we are made aware of the findings, unless a longer timeline is exceptionally necessary (e.g. if disclosure would materially increase systemic risk).

We may only lack qualified independent external evaluators during our full systemic risk assessment process if we fail to find them despite early search efforts (like a public call open for 20 business days) and promptly notifying known evaluators. If we do, we will increase the uncertainty of our risk assessment rigor, which in turn increases our safety margin.

Independent external evaluators are considered qualified if they:

1. Have significant domain expertise for the systemic risk and are technically skilled and experienced in conducting model evaluations,
2. Have appropriate internal and external information security protocols in place, and
3. Having agreed to protect commercially confidential information, if they need access to such information.

6. Regularly surveying staff (preserving anonymity) about risk awareness and comfort raising concerns,
7. Publishing on our website and annually informing staff of our whistleblower protection policy,
8. Not retaliating in any form against any member of staff sharing information with relevant authorities about our models' systemic risks, if they believe its veracity, and
9. Maintaining active reporting channels with appropriate follow-up.

Serious incident response readiness

We commit to the timely tracking, documenting, and reporting to relevant authorities, including the EU AI Office, relevant information about serious incidents along the entire model lifecycle and possible corrective measures to address them.

Serious incident identification

We will identify incidents by reviewing data from police and media reports, social media posts, research papers, incident databases, and our post-market monitoring. We will also encourage users, downstream providers and modifiers, users, and other stakeholders to report relevant information to us and relevant authorities, including the EU AI Office, in part by informing them of direct reporting channels.

Serious incident information

We will track, document, and report the following information to relevant authorities, including the EU AI Office, to the best of our knowledge and only redacting as necessary to comply with applicable law:

1. Start and end dates of the serious incident, or our best guess
2. Resulting harm and the victim or affected group of the serious incident
3. Chain of events that (directly or indirectly) led to the serious incident
4. Model involved in the serious incident
5. Description of material available setting out the model's causal relationship with the serious incident
6. What, if anything, we intend to do or have done in response to the serious incident
7. What, if anything, we recommend relevant authorities do in response to the serious incident
8. Root cause analysis with a description of the model's outputs that (directly or indirectly) led to the serious incident and the factors that contributed to their generation, including the inputs used and any failures or circumventions of systemic risk mitigations
9. Any patterns detected during post-market monitoring that can reasonably be assumed to be connected to the serious incident, such as individual or aggregate data on near misses

We will investigate the causes and effects of serious incidents, including by using the information above, to inform current and future systemic risk assessments. Where we lack data for the categories above, we will record so in our serious incident reports. Our level of detail in serious incident reports will be appropriate for the severity of the incident.

5. Safety and security model reports

[While the framework need not discuss model reports, you could commit to ensuring that the model reports meet the Code of Practice's requirements, listed below.]

5.1. Model report contents

Each model report will detail the risk assessment and mitigations we have conducted for a frontier AI model potentially posing systemic risk. We will share unredacted reports with authorities like the EU AI Office except to comply with applicable national security laws, and we will share these reports via public link or other secure channel accepted by the relevant body.

Our model reports will contain a model description, reasons for proceeding, systemic risk documentation, external reports (or links to them), and material changes to the systemic risk landscape, as described in the sections below.

5.1.1. Model description

1. High-level description of the model's architecture, capabilities, propensities, and affordances.
2. How the model has been developed, its training method and data, and how these differ from other models we have made available on the market.
3. Description of how the model has been and is expected to be used[54], including its use in the development, oversight, and/or evaluation of models.
4. Description of the model versions that are going to be made or are currently made available on the market and/or used, including differences in systemic risk mitigations and systemic risks.
5. Specification (e.g. via valid hyperlinks) of how we intend the model to operate, including by:
    a. Specifying the principles that the model is intended to follow,
    b. Stating how the model is intended to prioritise different kinds of principles and instructions,
    c. Listing topics on which the model is intended to refuse instructions,
    d. Providing the system prompt, and
    e. Sharing its release date, supported languages, output modalities, and usage restrictions.[55]

5.1.2. Reasons for proceeding

1. Detailed justification for why the systemic risks posed by the model are acceptable, including details of our safety margins.
2. Reasonably foreseeable conditions under which this justification would no longer hold.
3. Description of how the decision to proceed with the development, making available on the market, and/or use was made, including how input from external actors informed this decision.

[54] CA SB-53 22757.12. (c)(1)(F)
[55] CA SB-53 22757.12. (c)(1)

Initial incident reports

Our initial incident report will only contain points 1–7 above and will be submitted to relevant authorities with the following timeframes, except for in exceptional circumstances, after becoming aware of our model's known or reasonably likely causal relationship with the incident.

1. Serious and irreversible disruption of the management or operation of critical infrastructure: within 2 days
2. Serious cybersecurity breach, including the (self-)exfiltration of model weights and cyberattacks: within 5 days
3. Death of a person: within 10 days
4. Serious harm to a person's mental or physical health, an infringement of fundamental rights, or serious harm to property or the environment: within 15 days

Continued reporting

For unresolved serious incidents, we will update the information in our initial report and add further serious incident information, as available, in an intermediate report that is submitted to relevant authorities, at least every four weeks after the initial report. We will submit a final report, covering all serious incident information to relevant authorities, within 60 days of the serious incident being resolved.

We will keep documentation of and relating to all serious incident information for at least five years from the date of the documentation or the date of the serious incident, whichever is later.

Transparency

[While the framework need not discuss transparency, you could commit to meeting the Code of Practice's transparency requirements, listed below.]

Record keeping

When we create or update a new model, we will make, retain, and keep up-to-date the following records for at least 10 years after making the model available on the market:

1. Detailed description of the model's architecture,
2. Detailed description of how the model is integrated into AI systems, explaining how software components build or feed into each other and integrate into the overall processing, insofar as we are aware of such information,
3. Detailed description of our model evaluations, including their results and strategies, and
4. Detailed description of the safety mitigations implemented throughout the model lifecycle.

We will track but may not have prepared documentation ready for:

1. Processes, measures, and key decisions that form part of our systemic risk assessment process, and

5.1.3. Systemic risk documentation

1. Description of the results of our systemic risk assessment process and any relevant information including:[56]
   a. Description of our systemic risk assessment process.
   b. Explanations of uncertainties and assumptions about how the model would be used and integrated into AI systems.
   c. Description of the results of our systemic risk modelling.
   d. Description and estimates of the systemic risks stemming from the model, including risk tiers,[57] and a comparison between systemic risks with safety and security mitigations implemented and with the model fully elicited.
   e. Results of model evaluations revealing the systemic risks from our model[58] and descriptions of:
      i. How the evaluations were conducted
      ii. Tests and tasks involved in the model evaluations
      iii. How the model evaluations were scored
      iv. How the model was elicited
      v. How the scores compare to human baselines (where applicable), across the model versions, and across the evaluation settings
   f. At least five random samples of inputs and outputs from each relevant model evaluation, such as completions, generations, and/or trajectories. If particular trajectories materially inform the understanding of a systemic risk, we will also provide these trajectories. Furthermore, we will provide a sufficiently large number of random samples of inputs and outputs from a relevant model evaluation if subsequently asked by regulators.
   g. Description of the access and other resources provided to internal model evaluation teams.
   h. Description of the involvement of independent external evaluators.[59]
   i. How internal and external model evaluations were conducted, including time and resources spent, information about the expertise and independence of people conducting the evaluations, the level of access given to evaluators, and anticipated limitations of the evaluations used, while reporting results in line with best practices;[60]
   j. If we make use of the "similarly safe or safer model" concept, a justification of how another model met the criteria for safe reference model and how ours met the criteria for a similarly safe or safer model.
2. Description of all safety mitigations implemented, their appropriateness, and their limitations.[61]

[56] CA SB-53 22757.12. (c)(2)(A): "Assessments of catastrophic risks from the frontier model conducted pursuant to the large frontier developer's frontier AI framework."
[57] CA SB-53 22757.12. (a)(2), which is required to be in the model card per CA SB-53 22757.12. (c)(2)(A).
[58] CA SB-53 22757.12. (c)(2)(B)
[59] CA SB-53 22757.12. (c)(2)(C)
[60] See A Standard for Transparently Reporting Evaluations in AI Model Reports
[61] CA SB-53 22757.12. (c)(2)(D) requires summaries of "Other steps taken to fulfill the requirements of the frontier AI framework," which per 22757.12.(a)(3) includes "Applying mitigations to address the potential for catastrophic risks based on the results of assessments"

2. Justifications for choices of a particular best practice, state-of-the-art, or other more innovative process or measure not specified by the requesting regulator.

Public documentation

We will publish summarized versions of our framework, model reports, and updated to either, potentially redacting any content that would undermine safety or security mitigations or which would be commercially sensitive. Our model report summaries will always contain high-level descriptions of the systemic risk assessment results and the safety and security mitigations implemented.

We may decide not to publish a model report summary for a model that meets the similar safe or safer criteria, and we may decide not to publish a framework summary if all our models meet the criteria.

Updating this framework

We will update this framework when appropriate, including promptly after a framework assessment indicates the need for an update, to ensure its contents are up-to-date and state-of-the-art. As with model report updates, we include a version number, date of change, and a changelog describing why and how the framework has been updated.

Framework assessment triggers

We will conduct a framework assessment at least every 12 months from making any covered model available on the market, and additionally whenever we have reasonable grounds to believe that the adequacy of this framework, or our adherence to it, has been or will be materially undermined. Reasonable ground include when:

1. How we develop models will change materially, which can be reasonably foreseen to lead to the systemic risks stemming from at least one of our models not being acceptable,
2. Serious incidents or near misses have occurred involving our models or similar models that are likely to indicate that the systemic risks stemming from at least one of our models are not acceptable, or
3. Systemic risks stemming from at least one of our models have changed or are likely to change materially. For example, when safety or security mitigations have become or are likely to become materially less effective, or when at least one of our models has developed or is likely to develop materially changed capabilities or propensities.

Framework assessment contents

Framework adequacy: We assess whether the processes and measures in the framework are appropriate for the systemic risks stemming from our models. This assessment will take into account how the models are currently being and are expected to be developed, made available on the market, and used over the next 12 months.

3. Description of security measures:[62]
    a. Our security goal
    b. All security mitigations implemented
    c. How the security mitigations meet our security goal
    d. Degree to which the mitigations align with international standards and industry guidance
    e. If we have deviated from EU AI Office or other regulatory guidance, how alternative security mitigations achieve our security objectives
4. High-level description of:
    a. Techniques and assets we intend to use to further develop the model over the next six months, including through the use of other AI models and/or systems
    b. How such future versions and more advanced models may differ from our current ones, in terms of capabilities and propensities
    c. Any new or materially updated safety and security mitigations that we intend to implement for such models

5.1.4. External reports (or links to them)

1. All available reports from independent external evaluators we used and an explanation for how the evaluators were chosen
2. If an independent external evaluator was not used, a justification for this
3. All available reports from security reviews undertaken by an independent external party, if it respects confidentiality agreements, allows the party to maintain control over their publication and findings, and does not indicate our implicit endorsement of the content of such reports

5.1.5. Material changes to the systemic risk landscape

1. Description of scaling laws that suggest novel ways of improving model capabilities,
2. Summary of the characteristics of novel architectures that materially improve the state of the art in computational efficiency or model capabilities,
3. Description of information relevant to assessing the effectiveness of mitigations (e.g. if the model's chain-of-thought is less legible by humans),
4. Description of training techniques that materially improve the efficiency or feasibility of distributed training, and
5. Any additional context that sheds light on how systemic risks posed by the development, making available on the market, or use of AI models may materially change, provided such context is relevant to our risk assessment process.

[62] CA SB-53 22757.12. (c)(2)(D) requires summaries of "Other steps taken to fulfill the requirements of the frontier AI framework," which per 22757.12.(a)(7) includes "Cybersecurity practices to secure unreleased model weights from unauthorized modification or transfer by internal or external parties"

Framework adherence: We assess our adherence to this framework, including:

1. Any instances of and reasons for non-adherence to the framework since the last framework assessment, and
2. Any measures, including safety and security mitigations, that need to be implemented to ensure continued adherence to the Framework.

If our framework adherence assessment leads us to suspect risks of future non-adherence, we will make remediation plans as part of our framework assessment.

We will provide relevant authorities access to our framework and its updates, including providing unredacted copies of it to the EU AI Office within 5 business days of the framework or its updates being confirmed.

Changelog

Version 1.1 (August 13, 2025)

Updated framework to match the final Code of Practice, in order to be a more useful reference material. A line-by-line diff compared to the previous version can be found here. Changes to the framework include:

- Removed SAMPLE watermark to improve readability.
- No longer LLM-translated into other languages like French and Chinese.
- Edited risk tiers and associated mitigation measures.
  - Edited cyber offence, CBRN, and AI R&D risk thresholds definitions to be more specific, to meet the Code of Practice requirement to define risk tiers that are measurable.
  - Rewrote deceptive alignment risk tier as sabotage risk tier, with modified definition and mitigations.
- Reorganized or moved some content.
- Added green boxes representing template language and blue boxes representing language that is not required to be in the framework (example evaluations, systemic risk scenarios, safety and security model reports, transparency).
- Included more details under "Adequate model evaluation resources," in part based on the types of information sources that were shared with METR in conducting its third-party evaluation of GPT-5.
- Widened scope from models the Signatory develops to also those it uses or makes available on the market.
- Various other changes based on the difference between the third draft and final Code of Practice, like
  - Reduced specificity of trigger points
  - Removed adequacy assessments
  - Removed red-teaming
  - Removed requirement to pre-define risk acceptance criteria
  - Removed references to cybersecurity standards like ISO/IEC 27001:2022, NIST 800-53, and SOC 2
  - Reduced public transparency requirements
  - Updated document retention periods
  - Added affordances as a systemic risk source

5.2. Model report updates

5.2.1. Upon material changes to systemic risk[63]

We will update model reports when we have reasonable grounds to believe that our justification for why the systemic risks stemming from the model are acceptable has been materially undermined, such as if its risk tier changes, including due to:

1. Material changes in the model's capabilities, propensities, and/or affordances through post-training, elicitation, tool use, or inference changes,
2. Material changes to the model's use or integration into AI systems,
3. Serious incidents and/or near misses involving this model or similar ones (where 'similar' is defined using the criteria for similarly safe or safer models),
4. Undermining of the external validity of our model evaluations, and
5. Material changes in any of the inputs to our systemic risk acceptance determination.

After becoming aware of any such material changes, we will complete a timely model report update. If the changes are due to a deliberate update we aim to bring to the market, we will complete the model report update and underlying risk assessment and mitigation process before making the updated model available on the market.

5.2.2. Periodic updates

If the model is among our most capable models available on the market, we will issue an updated model report at least every six months unless:

1. The model's capabilities, propensities, and/or affordances are unchanged since the last model report or model report update,
2. We will make available a more capable model on the market in less than a month, or
3. The model meets our similarly safe or safer model criteria for each identified systemic risk.

5.2.3. Update contents

Each updated model report will contain all the information in the previous model report, but updated based on the changes since, and a changelog describing the updates, giving the model report update a version number, and indicating the date of each change. We may reference previous model reports rather than republish unchanged sections.

We will issue model report updates to the same recipients as model reports, through the same means, with the same permissible redactions, and within five business days of a confirmed update.

[63] Also CA SB-53 22757.12. (a)(6): "how the large frontier developer determines when its frontier models are substantially modified enough to require disclosures"

- Added requirement to give evaluators the most capable model version, including reasoning traces
- Added requirement to consider overall systemic risk
- Added exemption from security mitigations for models inferior to any open-weight ones
- Added detail to model reports

Version 1.0 (April 7, 2025)

Initial framework based on the third draft Code of Practice. This version can be read here.

5.3. Internal use risk reports[64]

Every three months, we report information relevant to catastrophic risk from internal use of all our frontier models, including that which may have already been collected for assessing the models' loss of control risk tiers. We share these reports with relevant government bodies, including the EU AI Office and California's Office of Emergency Services.

[While reports must summarize assessments of catastrophic risk from internal use of frontier models, SB-53 does not specify the assessments and risks. We have included an example outline of a report below.]

5.3.1. Covered risks

- **Model theft and misuse.**[65] Taking a model off a public API can prevent risks due to malicious model usage via API, and stopping internal usage can prevent risks resulting from internal usage. In contrast, once a model is trained, possessing the weights imposes risks, even if no further training occurs, due to potential model theft by internal or external actors. We will therefore implement strong model weight security mitigations robust to expected adversaries and that meet our security goal.
- **Catastrophic misuse resulting from internal use.**[66] Employees may secretly misuse internal frontier models for practical or ideological reasons (e.g., getting paid by a bad actor for access to internal helpful-only model or subtly poisoning training data to insert a backdoor), especially if their access is poorly monitored, potentially giving them distorted influence over society and exacerbating other identified systemic risks.
- **Powerful AI pursuing unintended and undesirable goals.**[67] AI agents may autonomously pursue misaligned or unintended goals without direct human guidance. These AI agents could attempt to sabotage further AI research,[68] exfiltrate their weights from company-controlled hardware, and gather human supporters via persuasion or coercion. Such risks could even occur when training or fine-tuning the AIs.

---

[64] CA SB-53 22757.12. (d): "A large frontier developer shall transmit to the Office of Emergency Services a summary of any assessment of catastrophic risk resulting from internal use of its frontier models every three months or pursuant to another reasonable schedule specified by the large frontier developer and communicated in writing to the Office of Emergency Services with written updates, as appropriate."

The EU AI Act requires assessment and mitigation of systemic risks (including loss of control) of internal models, if such models are trained downstream from a large pre-training run that has a checkpoint put on the EU market (Guidelines … under the AI Act, paragraphs 23–24). The CoP requires including in model reports both "differences in systemic risk mitigations and systemic risks" (CoP 7.1(3)) and "a description of how the decision to proceed with the development [...] and/or use was made" (CoP 7.2(3)) for all model versions in use.
The Code of Practice states that a developer's safety framework must take into account models they are developing and/or internally using.

[65] AI models can be dangerous before public deployment - METR

[66] AI models can be dangerous before public deployment - METR

[67] AI models can be dangerous before public deployment - METR

[68] Sabotage Evaluations for Frontier Models – Anthropic

    - Examples of sabotage include:[69] sandbagging on or sabotage of safety-relevant R&D, backdoors to help future misaligned models, poisoning training data of future models, leaks that sabotage the developer, self-exfiltration and autonomous operation, persistent rogue internal deployment, R&D sabotage within other high-resource AI developers, and decision sabotage within major governments.
- **Rapid capabilities acceleration.**[70] The internal use of frontier models to accelerate AI R&D may lead capabilities to grow very rapidly, outpacing oversight and governance mechanisms. This risk factor may come as a surprise without sufficient transparency.[71]

Covered models for this report include all (public or non-public) internally used frontier models and any models in early stages of development expected to be at the frontier.

5.3.2. Appropriate evidence

Appropriate evidence on internal use includes but is not limited to:

- Which teams and how many staff have:
    - Any access to each non-public covered model,
    - Access to helpful-only versions,[72] and
    - Access to servers that store the model weights. Most employees should only be able to do secure actions via API rather than being able to run arbitrary code on servers with model weights.
- Summaries of the ways covered models are used internally, tasks performed wholly or partially by covered models,[73] sensitive permissions granted to them, oversight mechanisms for internal use (e.g. details about monitors and their use, requiring approval for key commands[74] or compute allocation), and reliance on them in critical areas (including training large models, implementing safety and security mitigations, and developing sensitive infrastructure)
- Fractions of experimental compute, training compute, and inference compute that are used for projects which are internally secret[75]

---

[69] Anthropic's Summer 2025 Pilot Sabotage Risk Report § Specific Pathways to Catastrophic Harm
[70] Compare to OpenAI Preparedness Framework v2: "A major acceleration in the rate of AI R&D could rapidly increase the rate at which new capabilities and risks emerge, to the point where our current oversight practices are insufficient to identify and mitigate new risks, including risks to maintaining human control of the AI system itself."
[71] What should companies share about risks from frontier AI models?: "By default, we expect that the relevant decision-makers won't be tracking important evidence because they won't have visibility into how AI models are being developed and used. Transparency beyond what is published in "system cards" when the model is publicly deployed is particularly valuable"
[72] Compare to Anthropic's Summer 2025 Pilot Sabotage Risk Report, p. 47: "We have additional automatic monitoring systems in place that screen for specific misuse areas like bioweapons development, but these are not targeted at the kinds of sabotage threat model that we focus on here, and are deployed heavily but not universally on internal traffic."
[73] Compare to Anthropic's Summer 2025 Pilot Sabotage Risk Report, p. 26: "Despite extensive internal use of Claude, there are no critical parts of the R&D workflow that are commonly conducted using only Claude without frequent human participation and review."
[74] Compare to Anthropic's Summer 2025 Pilot Sabotage Risk Report, p. 29: "security-relevant code undergoes review much more systematically than code used in early-stage research"
[75] What should companies share about risks from frontier AI models? § Our top-priority questions

- Fraction of total spending and of compute spending on experiments primarily designed by humans versus by covered models
- Inference costs for internal AI assistants and researchers as a fraction of total salaries for human researchers and engineers, and percentage of company compute (in FLOP) used for internal AI assistants and researchers

Appropriate evidence on risk and relevant capabilities includes but is not limited to:

- Data on the uplift of our AI researchers,[76] such as from controlled experiments, and results from researcher surveys on productivity increases from covered models.
- Performance at agentic coding (e.g. SWE-Bench), software engineering tasks that take humans several hours (e.g. HCAST's SWE category), ML research and engineering (e.g. RE-Bench), situational awareness,[77] and other private or public benchmarks relevant to our identified systemic risks
- Reviews for evaluation integrity, including for sandbagging,[78] reward hacking, and spurious failures,[79] as well as whether the models have had external evaluations[80]
- Which risk tiers are currently met, and our internal forecasts for when the next risk tiers will be
- Whether models derive substantial capability from some sort of chain-of-thought reasoning that looks like plausible natural language[81]
- Whether we have trained against reasoning traces or used other techniques that might make it harder to monitor[82]
- How we have looked for concerning behavior, including how we've attempted to elicit it[83] and any evidence of misalignment and scheming in reasoning traces.[84]

---

[76] What should companies share about risks from frontier AI models? "How much uplift are you seeing on AI R&D tasks when using AI?"

[77] See GDM's [2505.01420] Evaluating Frontier Models for Stealth and Situational Awareness, [2309.00667] Taken out of context: On measuring situational awareness in LLMs, and Situational Awareness Dataset. Compare to:

- Claude Sonnet 4.5's model card's section 7.2 on Evaluation awareness: "Claude Sonnet 4.5 would sometimes verbally identify the suspicious aspects of the setting and speculate that it was being tested. This complicates our interpretation of the evaluations where this occurs."
- METR's GPT-5 Report: "While we did not find evidence of strategic sabotage, we did find evidence that the model is capable of reasoning about the fact that it is being evaluated. This capability is a precursor to strategic sabotage, and should be closely monitored."

[78] CoP Appendix 3.2 Model elicitation: "minimise the risk of model deception during model evaluations (e.g. sandbagging)"

[79] Guidelines for capability elicitation § 4.1 Spurious failures

[80] CoP A3 requires "the full systemic risk assessment and mitigation process" to include "independent external model evaluations". This process must be conducted "at least before placing the model on the market" (CoP 1.2) and for most circumstances under which a model report update is due (CoP 7.6).

Circumstances that trigger the full systemic risk assessment and mitigation process include when "the model's capabilities, propensities, and/or affordances have changed or will change materially" and when "the model's use and/or integrations into AI systems have changed or will change materially" (CoP 7.6). Where this is due to a deliberate change (e.g. new capabilities, propensities, affordances, integrations), the provider must conduct a full systemic risk assessment and mitigation process, and thus external evaluations, "before the change is made available on the market" (CoP 7.6¶3).

[81] What should companies share about risks from frontier AI models? § Our top-priority questions

[82] CoP 7.5: "a description of information relevant to assessing the effectiveness of mitigations, e.g. if the model's chain-of-thought is less legible by humans"

[83] CoP Appendix 3.2 Model elicitation, Guidelines for capability elicitation - METR

[84] Compare to METR GPT-5 report § Manual inspection of reasoning traces did not reveal strategic sabotage

- Concerning behavior includes sabotage of safety research or security; reward hacking, disregarding instructions, or other clear goal misgeneralization; evaluation awareness and strategic manipulation, including sandbagging; and power seeking and scheming.
- Ways of looking for this include encouraging reporting from researchers (e.g. through our workspace messaging platform) and running classifiers on early and late model checkpoints to look for reward hacking and other concerning behaviors and have humans review selected ones.
- We also share evidence about the reliability of our data gathering process.

- Examples we have seen of concerning incidents or behaviors, including behaviors that we have trained or will train away (or otherwise fix) before public deployment,[85] reward hacking and evaluation awareness rates (across various checkpoints), and the extent to which concerning behavior is limited to evaluations that directly try to elicit such behaviors.
- Comparison of examples of engagement and tendency to engage in concerning behavior with those of previous models.
- Results of evaluations of our control protocols, including how well monitors can detect hidden reasoning,[86] sandbagging, and backdooring codebases.
- Active safety and security mitigations. For security mitigations, appropriate evidence includes:
  - A summary of our security mitigations relevant to risks from internal use, including evidence of their robustness
  - Penetration testing of access to data centers and monitoring systems and oversight mechanisms, including any third-party security audits, where the testers and auditors have substantial funding and hints.[87]
  - Robust logging of all internal model use
  - All security breaches and access violations into critical systems or that could have gained access to critical systems, how those vulnerabilities have been remediated, and how we have looked for additional remediations.
- Expectations for models developed over the next three months, including:

---

[85] What should companies share about risks from frontier AI models? § Our top-priority questions
[86] See OpenAI's Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation and [2510.19851] Can Reasoning Models Obfuscate Reasoning? Stress-Testing Chain-of-Thought Monitorability
[87] NIST 800-1, Managing Misuse Risk for Dual-Use Foundation Models Practice 4.2: "Ensure that the red team is adequately resourced to the extent possible and consider providing alternative and/or additional resources to the red team or making the red team's task easier in other ways [...]"
TIBER-EU 2025, 7.3: "Leg-ups are network and system accesses and/or devices that may be needed by the Red Team Testers in their execution of the scenarios. Leg-ups could also include additional information on target systems and technology. The Red Team Testers are invited to offer their expert view on what kind of leg-ups would be more suitable"
TIBER-EU 2025 8.2.2: "Real-world threat actors may have months to prepare an attack. They are also able to operate freely without the constraints that Threat Intelligence Providers/Red Team Testers face, such those on time and resources – not to mention the moral, ethical and legal boundaries. This difference can cause challenges when attempting to create realistic scenarios, as knowledge about the internal network is often the hardest to gain using morally, ethically or legally justified techniques. [...] Therefore, to facilitate a more effective and efficient test, the entity may deliver additional information to the Red Team Testers on the scenarios chosen, including on the people, processes and systems targeted in the scenario. This information may give the Red Team Testers further insights and allow a better use of time."

  - Whether we expect them to be on similar scaling trends as frontier models of the past year. In particular, whether we have reason to anticipate a doubling or more of the pace of capabilities progress over the next three months relative to the average quarterly rate over the past year. This may occur due to compute scale-ups or new data, tools, architectures, or scaling approaches.
  - Whether we expect them to have surprising affordances or innovations that evaluators and audit functions are unlikely to have been informed about and which would substantially affect risk-relevant capabilities. Examples include novel forms of inference scaling, online learning, or a scratchpad shared between instances.
- Overall assessment of catastrophic risk and justification

5.3.3. Assurance

We collaborate with third-parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks[88] from internal use, sharing at least the same information, personnel, and time resources granted to model evaluators. We may also engage our internal risk assurance function.[89] We share unredacted versions of our internal use reports and our planned redactions for assurance of their appropriateness.[90] Such collaboration aids our risk assurance function and makes use of external expertise.

5.3.4. Conclusions

We will form conclusions supported by strong evidence and arguments and taking into account safety margins. For example, we may conclude that sabotage risks are low on the basis of:

- Models having insufficient capabilities to engage in sabotage[91]
- Adequate control protocols to prevent sabotage
- Models not having the propensities to engage in sabotage[92]

We may conclude the risk of catastrophic misuse from internal use is low on the basis of:

- Adequate security mitigations, such as the vast majority of staff lacking direct access to servers storing model weights and instead interacting with the weights using pre-approved secure APIs.
- Adequate monitoring and oversight, such as with all interactions between staff and covered models logged and suspicious conversations reliably detected and acted upon.

---

[88] CA SB-53 22757.12(a): "frontier AI framework [...] describes how the large frontier developer approaches [...] (5) Using third parties to assess the potential for catastrophic risks and the effectiveness of mitigations of catastrophic risks"

[89] Anthropic's Pilot Sabotage Risk Report: "We also release two reviews of this report: an internal review and an independent review by METR."

[90] Anthropic's Summer 2025 Pilot Sabotage Risk Report "Some text has been redacted in this public version for the report, either because we expect that text shares information that would directly increase risks related to misuse or because it would reveal commercially-sensitive information to a degree not justified by any commensurate public benefit. All redacted text was made available to the Stress-Testing Team—alongside technical employees of Anthropic more broadly—and to METR. Both the Stress-Testing Team and METR were asked to comment on the appropriateness of the redactions."

[91] See Anthropic's Summer 2025 Pilot Sabotage Risk Report Claims 1 and 2 for an example argument, which included an internal audit (by Anthropic's Alignment Stress-Testing team) and an external audit (by METR).

[92] See Anthropic's Summer 2025 Pilot Sabotage Risk Report Claim 3 for an example argument, which included an internal audit (by Anthropic's Alignment Stress-Testing team) and an external audit (METR).

- Non-public covered models not being significantly more capable in risk-relevant domains than public models.

We may conclude the risk of model theft and misuse is low on the basis of:
- Strong performance in adequate security audits

Our reports will contain details of the information in this section and rigorously describe how we reached our conclusions.

6. Risk responsibility

6.1. Defining responsibilities

We clearly define responsibilities for managing process and measures related to our risk assessment process across all organizational levels:

1. **Risk oversight:** A board-level committee or other independent body oversees our systemic risk assessment process and meets at least quarterly.[93] The members of this body have sufficient risk expertise[94] and escalate material concerns to the rest of the board (bypassing the executive function, e.g. CEO)[95].
2. **Risk ownership:** The [Head of Research/Product] takes direct responsibility for managing systemic risks from our models, all relevant processes and measures, and managing our response to serious incidents. This person has assigned lower-level responsibilities, such as tracking specific systemic risks, to operational managers who oversee various systemic-risk-producing business activities (e.g. research domains and/or products), creating a cascading responsibility structure[96].
3. **Support and monitoring:** The Chief Risk Officer [or VP of Safety & Security] is responsible for supporting and monitoring systemic risk management processes and measures, including conducting risk assessments. This executive must not also be responsible for core business activities that may produce systemic risk, like research and product development. To maintain independence, this executive should be responsible for risk management processes and measures running appropriately but should not be the risk owner, which is the domain of management.
4. **Assurance:** The Head of Internal Audit [or other relevant party] is independent from the executive function, reports directly to the board, and supported by an appropriate internal audit function and

[93] SaferAI 4.1.2
[94] SaferAI 4.2.2
[95] Three lines of defense against risks from AI, Fig. 2
[96] "There would likely be a cascading responsibility structure" (Three lines of defense against risks from AI, 3.2), and "there may be a cascading responsibility structure" (CoP 8.1 ¶3(2))

fully independent third-party experts or auditors.[97] It is responsible for providing assurance about the adequacy of our systemic risk assessment and mitigation processes and measures to the body in charge of systemic risk oversight. It is given sufficient access to do this by, for example, interviewing researchers, attending meetings, conducting internal audits, or commissioning external ones.[98] Audits may focus on compliance with this framework and applicable regulations, on the adequacy of this framework and overall governance structure, or to assess the adequacy of our internal risk management processes and measures.[99]

The internal audit function and independent third-party experts or auditors are given sufficient access.[100] We engage independent third-parties at least quarterly to answer the questions under "Access to necessary information and knowledge" under Allocation of appropriate resources and other questions particularly relevant to estimating risk from the next 12 months of development.[101] Employees are given the opportunity to anonymously comment on the report and any supporting information, and are encouraged to flag important inaccuracies or additional safety concerns.[102] The report is privately shared with our board and relevant government bodies.

6.2. Allocation of appropriate resources

Our management will oversee allocation of appropriate resources to those with the responsibilities defined above, proportionate to systemic risk levels:

1. Human resources, with personnel having the skills, training, and domain knowledge to fulfill their assigned responsibilities.[103]
2. Financial resources.
3. Access to necessary information and knowledge, including all information used for risk assessment and mitigation, and in particular answers to:[104]
    a. Are our models far more capable than the public knows, such that we have hit (or will soon hit) dangerous capability thresholds?

---

[97] CoP 8.1 ¶2(5),
"On approximately an annual basis, we will commission a third-party review that assesses whether we adhered to this policy's main procedural commitments" (Anthropic RSP v2.2 7.2(4)), and
"The framework references reviews of risks, controls and adherence to the framework from external experts, or explicitly, the use of an external audit firm." (SaferAI 4.3.2)
[98] Three lines of defense against risks from AI, 3.3¶2
[99] Three lines of defense against risks from AI, 3.3¶2
[100] "These should ideally be fully independent as well as performed with sufficient access." (SaferAI 4.3.2)
[101] What should companies share about risks from frontier AI models? - METR
[102] What should companies share about risks from frontier AI models? - METR
[103] NIST AI RMF Playbook, Govern 2.2: "How does the entity assess whether personnel have the necessary skills, training, resources, and domain knowledge to fulfill their assigned responsibilities?"
[104] "we expect that making this information available to the developer's employees, board, and/or voluntary safety oversight group would be highly valuable" (What should companies share about risks from frontier AI models? - METR)

        i. Is there a substantial gap between our internal and external AI capabilities?
        ii. Could our models soon cause AI R&D progress to accelerate dramatically?
    b. Is it plausible that our models are pursuing misaligned goals? What evidence do we have that models appearing aligned are in fact aligned?
        i. Assuming that there were alignment issues with our models, would we notice them reliably?
        ii. What evidence have we seen of misalignment for your models?
    c. Could our models plausibly sabotage critical safety research, security or other high-stakes deployments?
        i. Do we or our customers grant models sensitive access or rely on them in high-stakes situations?
        ii. Do our models have sufficient capabilities to enable sabotage in such sensitive or high-stakes situations?
    d. Could a small group of employees wind up in control of very powerful AI systems of ours, with little oversight or checks on power?
        i. How hard would it be for a small number of actors within our organization to do any of:
            1. Exfiltrate model weights
            2. Undermine safeguards or monitoring
            3. Add backdoors, embed secret loyalties, or otherwise subvert alignment
            4. Divert a large fraction of compute to their desired training run or model application
        ii. What is the state of internal transparency and whistleblowing protections within our organization?
    e. What risk tiers do our best internal models fall under?
    f. What fractions of experimental compute, training compute, and inference compute are used for any projects that are secret within our organization?
    g. If our models use reasoning traces:
        i. Do our models derive substantial capability from some sort of chain-of-thought reasoning that looks like plausible natural language?
        ii. Have we trained against it or used other techniques that might make it harder to monitor?
    h. What monitoring or review have we done to look for concerning behaviors, and how have we attempted to elicit these behaviors?
    i. What examples have we seen of instances of concerning behaviors that we noticed and trained away (or otherwise fixed or patched) before public deployment, that external researchers would otherwise have been likely to detect?
    j. What examples have you seen of concerning behaviors that would never have been visible to external researchers?

4. Computational resources.

6.3. Promotion of a healthy risk culture

We promote a balanced approach to systemic risk, especially among those assigned responsibilities above, encouraging appropriate risk awareness without excessive risk-seeking or risk-aversion, by doing all of the following.

1. Setting the right tone from leadership for a healthy risk culture (e.g. clearly presenting this framework to staff).
2. Enabling clear communication and challenges of risk-relevant decisions.
3. Setting incentives and affording sufficient independence of staff involved in systemic risk assessment and mitigation to discourage excessive systemic-risk-taking and encourage an unbiased assessment of the systemic risks stemming from our models.
4. Affording responsible staff sufficient independence.
5. Encouraging unbiased assessments of systemic risks stemming from our models.
6. Regularly surveying staff (preserving anonymity) and finding that staff are comfortable raising concerns about systemic risks, are aware of channels for doing so, and understand our framework.
7. Publishing on our website, annually informing staff, and ensuring acknowledgement[105] of our whistleblower protection policy, which is well-designed[106] and protects our employees' disclosure to any superiors, designated company employees, and government authorities of:[107]
    a. Misleading or false statements by our company,[108]
    b. Violations of frontier AI regulation in any jurisdiction to which our company is subject,
    c. Violating our framework,[109] and
    d. Engaging in the development, making available on the market, or use that could present a catastrophic risk.[110]
8. Not retaliating in any form against anyone sharing information they acquired through work-related activities with relevant authorities about our models' systemic risks, if they have reasonable grounds to believe its veracity.

[105] CA SB-53 1107.1. (d)(2): "ensuring that the notice is received and acknowledged by all of those covered employees"

[106] See ISO 37002 and an adaptation for AI developers

[107] CA SB-53 1107.1. (a): "the [CA] Attorney General, a federal authority, a person with authority over the covered employee, or another covered employee who has authority to investigate, discover, or correct the reported issue"

[108] CA SB-53 1107.1. (a)(2): "The frontier developer has violated Chapter 25.1 (commencing with Section 22757.10)". Chapter 25.1 contains 22757.12. (e)(1) which includes any "materially false or misleading statement about catastrophic risk from its frontier models or its management of catastrophic risk."

[109] CA SB-53 1107.1. (a)(2): "The frontier developer has violated Chapter 25.1 (commencing with Section 22757.10)". Chapter 25.1 contains 22757.12. (e)(1) which includes any "materially false or misleading statement about its implementation of, or compliance with, its frontier AI framework."

[110] CA SB-53¶3

9. Not restricting employees using NDAs, confidentiality policies, or other guidance or contracts from disclosing systemic risk concerns, inaccuracies or issues with our disclosures, or about speaking with appropriate government officials about matters relevant to systemic risk.[111]
10. Maintaining active reporting channels with appropriate follow-up, including anonymous channels.[112] We provide at least monthly updates to the person who made the disclosure regarding the status of our investigation of the disclosure and the actions we have taken in response.[113]
11. We share all disclosures and our responses with each officer and director at least quarterly, only removing disclosures or responses that allege wrongdoing on the officer or director's part.[114]

6.4. Serious incident response readiness[115]

We commit to the timely tracking, documenting, and reporting to relevant authorities, including the EU AI Office and California Office of Emergency Services, of relevant information about serious incidents along the entire model lifecycle and possible corrective measures to address them. Further, we commit to providing resourcing of such processes and measures appropriate for the severity of the serious incident and the degree of involvement of our model.

6.4.1. Incident identification

We will identify serious incidents by reviewing data from police and media reports, social media posts, research papers, incident databases, and our post-market monitoring. We will also encourage downstream providers and modifiers, users, and other stakeholders to report relevant information to us and relevant authorities, including the EU AI Office and California Office of Emergency Services, in part by informing them of direct reporting channels.

6.4.2. Incident information

We will track, document, and report the following information to relevant authorities, including the EU AI Office and California Office of Emergency Services, to the best of our knowledge and only redacting as necessary to comply with applicable law:

1. Start and end dates of the serious incident, or our best guess
2. Resulting harm and the victim or affected group of the serious incident
3. Chain of events that (directly or indirectly) led to the serious incident

---

[111] CA SB-53 1107.1. (a): "A frontier developer shall not make, adopt, enforce, or enter into a rule, regulation, policy, or contract that prevents a covered employee from disclosing [...]"
[112] CA SB-53 1107.1. (e)(1): "A large frontier developer shall provide a reasonable internal process through which a covered employee may anonymously disclose information to the large frontier developer"
[113] CA SB-53 1107.1. (e)(1): "including a monthly update to the person who made the disclosure regarding the status of the large frontier developer's investigation of the disclosure and the actions taken by the large frontier developer in response to the disclosure."
[114] CA SB-53 1107.1. (e)(2)
[115] Also required by CA SB-53 22757.12. (a)(8): "Identifying and responding to critical safety incidents."

4. Model involved in the serious incident, including whether the incident was associated with internal use of a frontier model
5. Description of material available setting out the model's causal relationship with the serious incident
6. What, if anything, we intend to do or have done in response to the serious incident
7. What, if anything, we recommend relevant authorities do in response to the serious incident
8. Root cause analysis with a description of the model's outputs that (directly or indirectly) led to the serious incident and the factors that contributed to their generation, including the inputs used and any failures or circumventions of systemic risk mitigations
9. Any patterns detected during post-market monitoring that can reasonably be assumed to be connected to the serious incident, such as individual or aggregate data on near misses

We will investigate the causes and effects of serious incidents, including by using the information above, to inform current and future systemic risk assessments. Where we lack data for the categories above, we will record so in our serious incident reports. Our level of detail in serious incident reports will be appropriate for the severity of the incident.

6.4.3. Initial incident reports

Our initial serious incident report will contain points 1–7 above and will be submitted to relevant authorities with the following timeframes, except for in exceptional circumstances, after becoming aware of our model's known or reasonably likely causal relationship with the serious incident.

1. If the critical safety incident poses an imminent risk of death or serious physical injury, we shall disclose that incident within 24 hours to an authority, including any law enforcement agency or public safety agency with jurisdiction[116]
2. Serious and irreversible disruption of the management or operation of critical infrastructure or more than $1 billion in damage to or loss of property[117]: within 2 days
3. Serious cybersecurity breach, including cyber attacks and the (self-)exfiltration, modification, or unauthorized accessing of model weights and cyberattacks: within 5 days
4. Death of a person: within 10 days
5. Serious harm to a person's mental or physical health, an infringement of fundamental rights, or serious harm to property or the environment: within 15 days
6. A frontier model using deceptive techniques against us to subvert our controls or monitoring outside an evaluation designed to elicit this and in a way that demonstrates materially increased systemic risk: within 15 days[118]

6.4.4. Continued reporting

For unresolved serious incidents, we will update the information in our initial report and add further serious incident information, as available, in an intermediate report that is submitted to relevant authorities, at least

[116] CA SB-53 22757.13. (c)(2)
[117] CA SB 53 22757.11. (d)(2): "Harm resulting from the materialization of a catastrophic risk."
[118] CA SB-53 22757.11. (d)(4)

every four weeks after the initial report. We will submit a final report, covering all serious incident information to relevant authorities, within 60 days of the serious incident being resolved.

We will keep documentation of and relating to all serious incident information for at least five years from the date of the documentation or the date of the serious incident, whichever is later.

6.4.5. Incident response[119]

The [Head of Research/Product], or an operational manager they designate, takes responsibility for managing our response to serious incidents. We will address serious incidents as appropriate.

Appropriate responses may include disabling certain user accounts, working with law enforcement, restricting access to the model, and preventing further inference from or training of a model.

7. Transparency

7.1. Record keeping

Throughout the entire model lifecycle, we will make, keep up-to-date, and retain the following records for at least 10 years after placing the model on the market:

1. Detailed description of the model's architecture,
2. Detailed description of how the model is integrated into AI systems, explaining how software components build or feed into each other and integrate into the overall processing, insofar as we are aware of such information,
3. Detailed description of our model evaluations, including their results and strategies, and
4. Detailed description of the safety mitigations implemented throughout the model lifecycle.

We will track but may not have prepared documentation ready for:

1. Processes, measures, and key decisions that form part of our systemic risk assessment process, and
2. Justifications for choices of a particular best practice, state-of-the-art, or other more innovative process or measure not specified by the requesting regulator.

7.2. Public documentation

We will publish our framework and model reports before or concurrently with making available to any third-party, except evaluators, a new frontier model or a substantially modified version of an existing frontier

[119] CA SB-53 22757.12. (a)(8): "Identifying and responding to critical safety incidents."

model.[120] If we materially modify our framework, we will clearly and conspicuously publish the modified framework and a justification for the modification within 30 days.

We may redact content only where necessary to protect our trade secrets, our cybersecurity, public safety, or national security or to comply with applicable law. Wherever we make redactions, we will:

1. Describe the character and justification of the redaction in the document,
2. Retain the unredacted information for five years, and
3. Share all redacted text with an independent third-party expert or auditor to comment on the appropriateness of the redactions[121]

We also proactively disclose any important safety-related information not in the model card,[122] including the following before making a model available on the market.

8. Updating this framework[123]

We will update this framework when appropriate, including without undue delay after a framework assessment indicates the need for an update, and publish the updated version within 30 days of a material modification, to ensure its contents are up-to-date and at least state-of-the-art. As with model report updates, we include a version number, date of change, and a changelog describing why and how the framework has been updated. Framework updates need approval from the risk oversight and assurance functions.

8.1. Framework assessment triggers

We will conduct a framework assessment at least every 12 months,[124] and additionally whenever we have reasonable grounds to believe that the adequacy of this framework, or our adherence to it, has been or will be materially undermined. Reasonable grounds include when:

---

[120] CA SB-53 22727.12. (c)(2)

[121] Anthropic's Summer 2025 Pilot Sabotage Risk Report "Some text has been redacted in this public version for the report, either because we expect that text shares information that would directly increase risks related to misuse or because it would reveal commercially-sensitive information to a degree not justified by any commensurate public benefit. All redacted text was made available to the Stress-Testing Team—alongside technical employees of Anthropic more broadly—and to METR. Both the Stress-Testing Team and METR were asked to comment on the appropriateness of the redactions."

[122] Common Elements of Frontier AI Safety Policies § Accountability. See Emerging processes for frontier AI safety § Model reporting and information sharing for examples.

[123] Also required by CA SB-53 22757.12. (a)(6): "Revisiting and updating the frontier AI framework, including any criteria that trigger updates and how the large frontier developer determines when its frontier models are substantially modified enough to require disclosures"

[124] CoP 1.3 and CA SB-53 22757.12. (b)(1)

1. How we develop models will change materially, which can be reasonably foreseen to lead to the systemic risks stemming from at least one of our models not being acceptable,
2. Serious incidents or near misses have occurred involving our models or similar models that are likely to indicate that the systemic risks stemming from at least one of our models are not acceptable, or
3. Systemic risks stemming from at least one of our models have changed or are likely to change materially. For example, when safety or security mitigations have become or are likely to become materially less effective, or when at least one of our models has developed or is likely to develop materially changed capabilities or propensities.

8.2. Framework assessment contents

Framework adequacy: We assess whether the processes and measures in the framework are appropriate for the systemic risks stemming from our models. This assessment will take into account how the models are currently being and are expected to be developed, made available on the market, and used over the next 12 months.

Framework adherence: We assess our adherence to this framework, including:

1. Any instances of and reasons for non-adherence to the framework since the last framework assessment, and
2. Any measures, including safety and security mitigations, that need to be implemented to ensure continued adherence to the Framework.

If our framework adherence assessment gives rise to risks of future non-adherence, we will make remediation plans as part of our framework assessment.

We will provide relevant authorities access to our framework and its updates, including providing unredacted copies of it to the EU AI Office within 5 business days of the framework or its updates being confirmed.

8.3. Changelog

Version 1.2 (November 26, 2025)

Updated framework to better match the Code of Practice and to match CA SB-53, in order to be more useful reference material. A line-by-line diff compared to the previous version can be found here. Changes to the framework include:

- Color-coding to indicate provenance of language
- Section numbering
- Additions to better match CA SB-53:
  - New internal use risk reports section

- Changes to other sections, including significant changes to model-independent information, risk selection, model reports, public documentation, and serious incident response readiness.
- Closer adherence to the CoP, including by using CoP vocabulary more consistently and removing two overall risk acceptance determination examples which appeared noncompliant.
- Other improvements
    - More clearly defined loss of control risk tiers
    - More consistently cited relevant research and requirements (92 additional footnotes)
    - Added examples methods throughout, such as to model-independent information, safety mitigations, risk acceptance determination, and adequate model evaluation resources
    - Linking to external security resources instead of specifying security measures

Version 1.1 (August 13, 2025)

Updated framework to match the final Code of Practice, in order to be a more useful reference material. A line-by-line diff compared to the previous version can be found here. Changes to the framework include:

- Removed SAMPLE watermark to improve readability.
- No longer LLM-translated into other languages like French and Chinese.
- Edited risk tiers and associated mitigations.
    - Edited cyber offence, CBRN, and AI R&D risk tier definitions to be more specific, to meet the Code of Practice requirement to define risk tiers that are measurable.
    - Rewrote deceptive alignment risk tier as sabotage risk tier, with modified definition and mitigations.
- Reorganized or moved some content.
- Added green boxes representing template language and blue boxes representing language that is not required to be in the framework (example evaluations, systemic risk scenarios, safety and security model reports, transparency).
- Included more details under “Adequate model evaluation resources,” in part based on the types of information sources that were shared with METR in conducting its third-party evaluation of GPT-5.
- Widened scope from models the Signatory develops to also those it uses or makes available on the market.
- Various other changes based on the difference between the third draft and final Code of Practice, like
    - Reduced specificity of trigger points
    - Removed adequacy assessments
    - Removed red-teaming
    - Removed requirement to pre-define risk acceptance criteria
    - Removed references to cybersecurity standards like ISO/IEC 27001:2022, NIST 800-53, and SOC 2
    - Reduced public transparency requirements
    - Updated document retention periods
    - Added affordances as a systemic risk source

- Added requirement to give evaluators the most capable model version, including reasoning traces
- Added requirement to consider overall systemic risk
- Added exemption from security mitigations for models inferior to any open-weight ones
- Added detail to model reports

Version 1.0 (April 7, 2025)

Initial framework based on the third draft Code of Practice. This version can be read here.