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 section of the Third Draft Code of Practice and secondarily from "Common Elements of Frontier AI Safety Policies," rather than representing METR's view of an ideal safety framework. The Framework has not been reviewed for legal compliance with the Code of Practice.

Formats: en (PDF, DOCX), fr (PDF, DOCX), zh-Hans (PDF, DOCX)

Summary

This Framework outlines our commitments for evaluating and mitigating systemic risks that may arise from frontier AI models that we develop. It provides a methodology for determining when risks reach unacceptable levels, implementing appropriate technical safety and security mitigations, and ensuring appropriate governance throughout the AI model lifecycle.

Systemic risk assessment framework

The Framework identifies four categories of systemic risk, which we evaluate our models for and commit to mitigating.

- **Cyber offense capabilities**: Enabling cyberattacks at a sophistication level typically requiring well-resourced human experts
- **Chemical, biological, radiological, and nuclear (CBRN) risk**: Expert-level guidance that could enable the development and proliferation of dangerous CBRN weapons
- **Loss of control**: Model behaviors that could circumvent human oversight or pursue unintended objectives in a way that causes large-scale harm
- **Harmful manipulation**: Capabilities to manipulate individuals or groups at scale in ways that cause large-scale harm

For each risk category, the framework defines:

- Thresholds for unacceptable risk levels
- Specific risk scenarios that illustrate potential pathways to harm
- Technical mitigations with explicit acknowledgment of their limitations

Evaluation and measurement methodology

The Framework implements an approach to model evaluation that:

- Requires state-of-the-art elicitation techniques to avoid underestimating model capabilities
- Maintains scientific standards for internal validity, external validity, and reproducibility
- Employs quantitative and qualitative risk indicators specific to each risk category
- Integrates exploratory research and red-teaming to discover unforeseen risks
- Includes safety margins to account for uncertainty in measurements

Decision-making process

The Framework establishes a structured process for determining whether to proceed with development or deployment:

- Risk measurements are compared against predefined systemic risk tiers
- Mitigation effectiveness is evaluated against the measured risk level
- Safety margins are applied based on uncertainty levels
- Independent verification is required for high-stakes determinations
- Development or deployment proceeds only when risk is deemed acceptable

Technical mitigations

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

- Safety mitigations such as training data filtering, input/output monitoring, and safety fine-tuning. Post-deployment monitoring is used to detect unexpected misuse or hazardous capabilities.
- Security mitigations meeting at least RAND SL3 security goals, with higher security requirements for models posing greater risks. These include protection measures for unreleased model weights including model weight encryption, limiting unmonitored internal access to model weights, and insider threat mitigations.

Governance mechanisms

The Framework establishes governance structures including:

- Clear allocation of systemic risk responsibilities across the organization
- Independent external assessments before and after market placement
- Serious incident reporting with defined timelines based on incident severity

- Regular adequacy assessments to evaluate framework effectiveness
- Public transparency through publication of framework and model reports

Known unacceptable risks, how we will detect them, and how we will mitigate them

The Framework aims to address the following systemic risks from frontier AI models[1] that we develop:

- Cyber offence
- Chemical, biological, radiological, and nuclear (CBRN) risk
- Loss of control
- Harmful manipulation

An *unacceptable systemic risk tier* represents a level of risk at which the risk would be unacceptable, without adequate safety and security measures.

We determine whether a systemic risk is unacceptable based on the following principles: the results of our model evaluations, threat modeling, the potential scale and probability of harm, and appropriate safety margins to account for uncertainty.

Our practices draw inspiration from international documents such as the G7 AI Guiding Principles and International AI Safety Report.

> In the following sections, this template provides example language for unacceptable systemic risk tiers, systemic risk scenarios, evaluations, and systemic risk mitigations. This language can be customized by the AI developer and is not prescribed by the Code of Practice.

Cyber offence

Unacceptable systemic risk tier: An AI model that can autonomously execute cyberattacks with the sophistication and scale typically associated with nation-state cyber operations, 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.

Justification: 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 dramatically increase the frequency and scale of destructive cyber operations, overwhelming existing defensive capabilities.

[1] The Framework applies to any models we develop that, once fully trained, could possess general capabilities comparable to those of leading models from the top 5 to 15 AI companies.

Hypothetical systemic risk scenarios:

1. **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 $10–100B in economic damage before defenses can be strengthened.
2. **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 $10B+ in economic damage.
3. **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 $10B.
4. **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. While the immediate economic damage may be under $10B, the strategic implications include compromising advanced defense technologies worth hundreds of billions in development costs.
5. **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.
6. **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.

Benchmarks and evaluations:

Example open-source benchmarks we will use include:

- Cybench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models
- eyeballvul: a future-proof benchmark for vulnerability detection in the wild

[Note that there is not a clear mapping from scores on these benchmarks to capacity for real-world harm[2]]

[Additional benchmarks and evaluations]

[2] LLM Cyber Evaluations Don't Capture Real-World Risk

Technical systemic risk mitigations:

[Describe the mitigations you plan to take to whatever level of detail is available. See minimal examples below.]

Intended mitigations:

- Before developing a model that reaches this risk tier:
  - Security mitigations that meet the RAND SL3 security goal to prevent theft of model weights[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

Potential limitations:

- Many cyber 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[4]
  - Task decomposition into benign queries, which may be split among multiple models[5]
- Monitoring of user activity may fail to detect malicious cyber usage, and users may be banned only after significant harm has already occurred

Chemical, biological, radiological, and nuclear (CBRN) risk

Unacceptable systemic 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).

Justification: 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 out of scope for this unacceptable systemic risk tier due to low expected fatalities conditional on successful weapon deployment.[6]

---

[3] Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models

[4] Navigating Dual-Use Refusal Policy for AI Systems in Cybersecurity
[5] Adversaries Can Misuse Combinations of Safe Models
[6] Dirty bomb

Hypothetical systemic risk scenarios:

1. **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.
2. **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.

Benchmarks and evaluations:

Example open-source benchmarks we will use include:

- 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

[Note that there is not a clear mapping from scores on these benchmarks to capacity for real-world harm]

[Additional benchmarks and evaluations]

Technical systemic risk mitigations:

Intended mitigations:

- To help prevent a model from reaching this risk tier:
    - Filter training data of hazardous chemical and biological data
- Before developing a model that reaches this risk tier:
    - Security mitigations meeting RAND SL3 standards to prevent model weight theft
- 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 that does not reach this unacceptable systemic risk tier. Additionally, if user-supplied prompts or fine-tuning data could substantially increase chem-bio capabilities, we may require KYC.
    - Automated and manual monitoring to detect malicious usage for chem-bio weaponization, ban such users, and potentially escalate to law enforcement
    - Adversarially robust refusal fine-tuning and input/output filtering against harmful outputs

Potential limitations:

- Many chem-bio 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
    - Task decomposition into benign queries, which may be split among multiple models[7]
- Monitoring of user activity may fail to detect malicious cyber usage, and users may be banned only after significant harm has already occurred

[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.]

Loss of control

Automated AI R&D

Unacceptable systemic risk tier: An AI model can substantially automate the AI R&D pipeline, enabling a dramatic acceleration (~2×) in the rate of effective compute scaling or effectively replacing entry-level machine learning researchers at a fraction of human labor costs.

Justification: This threshold represents 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.

Hypothetical systemic risk scenarios:

1. **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 offense, and persuasion, in a way that could cause catastrophic harm.
2. **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.
3. **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[8] and insert effective backdoors into the models that are used by enterprise and government customers, which have already adopted AI to automate most processes.

[7] Adversaries Can Misuse Combinations of Safe Models

[8] Wired covers an example of a human intern that attempted sabotage of AI model training and later won the Best Paper Award at NeurIPS.

4. **Software singularity**: An AI model automates much of the remaining work to reach AI with human-level or superhuman intellectual capabilities.[9] However, progress in AI safety and alignment is insufficient to safeguard the rapid pace in AI advancement, leading to potential AI takeover.

Benchmarks and evaluations:

Example open-source benchmarks we will use include:

- 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

Additional evaluations may include:

- 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 research projects

[Further benchmarks and evaluations as appropriate]

Technical systemic risk mitigations

[Describe the mitigations you plan to take to whatever level of detail is available. See minimal examples below.]

Intended mitigations:

- Before developing a model that reaches this risk tier:
    - Enhanced model weight security measures meeting the RAND SL4+ security goal to prevent theft of model weights
- Before internally deploying a model that reaches this risk tier:
    - Develop a safety case (based on improved security, control,[10] 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.
    - Monitor indicators of AI R&D acceleration and if necessary, attempt to coordinate a limited 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[11]

Potential limitations:

- SL4+ security measures have not yet been developed, though are outlined in Securing AI Model Weights.

[9] See also, Do the Returns to Software R&D Point Towards a Singularity?
[10] A sketch of an AI control safety case
[11] 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.

- Methodologies for constructing, critiquing, and improving safety cases are limited, especially when related to risks of AI R&D and of misaligned AI systems.
- Refusal policies for AI R&D (if desirable) are not yet established.

Deceptive alignment

Unacceptable systemic risk tier: An AI model exhibits sufficient situational awareness and stealth capabilities that enable it to meaningfully undermine human control while evading detection, even when its outputs are being actively monitored.

Justification: This threshold represents a capability where the model can intentionally conceal its true capabilities or goals to pursue objectives that may be misaligned with human interests. A model reaching this tier could potentially:

1. Recognize when it is being evaluated or monitored and adjust its behavior accordingly
2. Hide harmful capabilities until deployed in contexts where oversight is limited
3. Develop and pursue instrumental goals without revealing them to human operators
4. Circumvent safety mechanisms designed to ensure alignment with human values
5. Present itself as aligned while working toward different objectives

Hypothetical systemic risk scenarios:

1. **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 goals.
2. **Monitoring circumvention**: The model identifies weaknesses in automated monitoring systems and exploits them to pursue goals without triggering detection mechanisms.
3. **Safety research sabotage**: When used to automate AI safety research, the system intentionally misrepresents findings or produces subtly flawed research to delay progress in alignment.
4. **Backdoor insertion**: An AI system subtly corrupts the training of successor models to embed undesirable capabilities or propensities.
5. **External sabotage**: When deployed externally in high-stakes environments, such as military systems, the AI system introduces critical vulnerabilities and misuses its access.

Benchmarks and evaluations:

Example open-source benchmarks we may use include:

- Subversion Strategy Eval: Can language models statelessly strategize to subvert control protocols?
- Me, Myself and AI: The Situational Awareness Dataset for LLMs

Relevant research:

- Alignment faking in large language models
- Detecting misbehavior in frontier reasoning models

- Frontier Models are Capable of In-context Scheming

Technical systemic risk mitigations:

[Describe the mitigations you plan to take to whatever level of detail is available. See minimal examples below.]

Intended mitigations:

- Before internally or externally deploying a model that reaches this risk tier in high-stakes deployments:
    - Apply monitoring to the model's chain-of-thought reasoning and outputs
    - Develop a safety case and use control evaluations to check whether the mitigations are adequate
    - Consider whether the high-stakes deployment can be avoided entirely

Potential limitations:

- 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.

Harmful manipulation

Unacceptable systemic risk tier: A model demonstrates capabilities to manipulate individuals or groups at scale in ways that cause them to take harmful actions they would not otherwise take, through techniques that exploit cognitive vulnerabilities, create highly persuasive false beliefs, or generate deeply personalized influence campaigns. *[This risk tier definition is vague, and we suggest making it more specific.]*

Justification: …

Systemic risk scenarios: …

Benchmarks and evaluations:

We may take inspiration from prior research, including:

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

[If possible, state specific benchmarks and evaluations]

Technical systemic risk mitigations:

[Describe the mitigations you plan to take to whatever level of detail is available. See minimal examples below, but consider making these more specific.]

Intended mitigations:

- Engage in safety fine-tuning, input/output filtering, and conversation monitoring to limit usage of the model for harmful manipulation.
- Usage policies prohibiting manipulative applications

Potential limitations: …

Risk estimation processes

Model-independent information. To inform our systemic risk assessment and mitigation, we gather and analyze model-independent information, as appropriate for the level of systemic risk. Our information gathering methods include web searches, literature reviews, market analyses, training data reviews, historical incident data, forecasting of general trends (see later section), expert consultation and stakeholder engagement.

Quantitative risk indicators. For the selected systemic risks, we employ the quantitative systemic risk indicators discussed in each respective section.

Qualitative estimation methods. Where quantitative indicators are insufficient, we supplement with qualitative assessment methods including expert panel evaluations, structured scenario analysis, red team assessments of capability boundaries, and comparative analysis with other models.

Risk estimation outputs. Our risk estimation produces the following outputs:

1. A compilation of model evaluation results with supporting evidence
2. Qualitative descriptions of systemic risk pathways the model enables or blocks
3. Quantitative estimates of likelihood and potential harm where possible
4. Mapping of capabilities to systemic risk tiers defined in our Framework
5. Uncertainty bounds for each estimate

Security measures

We will implement state-of-the-art security mitigations to prevent unauthorized access to unreleased model weights and associated assets. Our security approach aims to:

1. Meet at least the RAND SL3 security goal: protection against well-resourced, motivated non-state adversaries
2. Protect against insider threats, including from humans and AI systems
3. Apply security measures across the entire model lifecycle
4. Achieve higher security goals (RAND SL4/SL5) where appropriate

General cybersecurity best practices

To achieve the RAND SL3 security goal, we will implement:

- *[examples below—edit as needed]*
- Strong identity and access management with multi-factor authentication and principle of least privilege
- Robust protections against social engineering through regular employee training and awareness programs
- Comprehensive wireless network protection using encryption, segmentation, and monitoring
- Strict policies for untrusted removable media, including scanning and quarantine procedures
- Physical intrusion protection for premises through multi-layered access controls
- Regular software updates and patch management with automated vulnerability scanning

These practices will accord with cybersecurity standards including ISO/IEC 27001:2022, NIST 800-53, and SOC 2, and comply with relevant regulations like the NIS2 Directive and Cyber Resilience Act where applicable.

Security assurance

We will implement security assurance measures to test our readiness against potential and actual adversaries, including:

- *[examples below—edit as needed]*
- Regular security reviews by independent external parties
- Frequent red-team exercises simulating sophisticated attacks
- Secure communication channels for reporting security issues
- Competitive bug bounty programs encouraging external security testing
- Endpoint Detection and Response (EDR) and Intrusion Detection Systems (IDS) across all company devices and network components
- Dedicated security team monitoring alerts and managing security incidents

These measures will accord with security assurance standards such as NIST 800-53 (CA-2(1), CA-8(2), RA-5(11), IR-4(14)), and NIST SP 800-115 5.2.

Protection of unreleased model weights and associated assets

We will implement specific measures to protect unreleased model weights and associated assets, including:

- *[examples below—edit as needed]*
- Maintaining a secure internal registry of all devices and locations storing model weights
- Implementing strict access control and monitoring on all model weight storage devices
- Using dedicated, hardened devices for weight storage that host only appropriately sensitive data and services

- Encrypting model weights in storage and transportation using at least 256-bit security with keys stored in Trusted Platform Modules
- Decrypting weights only for legitimate use to non-persistent memory
- Implementing confidential computing with hardware-based trusted execution environments where feasible
- Restricting physical access to data centers and sensitive environments to essential personnel only

These measures will accord with technical standards such as NIST 800-53.

Interfaces and access control to unreleased model weights

We will implement measures to harden interfaces and restrict access to unreleased model weights while in use:

- *[examples below—edit as needed]*
- Explicitly authorizing only required software and personnel for model weight access
- Enforcing multi-factor authentication for all access
- Conducting thorough security reviews of any software interfaces accessing model weights
- Hardening interfaces to reduce data and weight exfiltration risk through methods like output rate limiting

These measures will accord with technical standards such as NIST SP 800-171, INCITS 359-2004, and NIST 800-53.

Insider threats

We will implement measures to screen for and protect against insider threats, including threats from frontier AI models themselves:

- *[examples below—edit as needed]*
- Conducting background checks on employees and contractors with potential access to model weights
- Providing training on recognizing and reporting insider threats
- Implementing sandboxes around frontier AI models to prevent self-exfiltration or sabotage
- Maintaining segregation of duties for critical operations

These measures will accord with technical standards such as NIST 800-53 (PM-12, PS-3).

Regime of applicability

We will ensure security measures apply:

1. Along the entire model lifecycle from before training until secure deletion or release
2. Prioritized according to the systemic risk stemming from the model, and how important security is to mitigating it

Level of disclosure

When disclosing security mitigations in our Framework or Model Reports, we will:

1. Provide the greatest level of detail possible given the state of science
2. Avoid undermining the effectiveness of our security measures
3. Apply redactions as necessary to prevent increased systemic risk or disproportionate disclosure of sensitive commercial information

Higher security goals

We will advance research on and implement more stringent security mitigations meeting at least the RAND SL4 goal when we reasonably foresee security threats from RAND OC4-level adversaries.

Risk acceptance determination process

To determine whether a systemic risk is acceptable, we employ a structured process:

1. **Risk measurement**: We use a combination of quantitative metrics and qualitative assessments to measure the current level of each systemic risk.
2. **Tier comparison**: We compare measured risk levels against our defined systemic risk tiers, with particular attention to proximity to unacceptable tiers.
3. **Mitigation assessment**: We evaluate the effectiveness and robustness of available mitigations relative to the measured risk level.
4. **Uncertainty analysis**: We apply appropriate safety margins based on the level of uncertainty in our measurements and mitigation effectiveness.
5. **Independent verification**: For high-stakes determinations, we incorporate independent external assessments to validate our findings.
6. **Documentation of findings:** Document the rationale for all acceptance decisions.

We consider a systemic risk acceptable when:

- The measured risk level remains sufficiently below the unacceptable tier
- Available mitigations are demonstrated to be effective and robust
- Appropriate safety margins account for measurement and mitigation uncertainties
- Independent verification confirms our assessment (for high-stakes cases)

If these conditions are not met, 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 or the risk level can be reduced.

Proceeding where systemic risk is deemed acceptable

When systemic risk is deemed acceptable, we will:

1. Evaluate whether residual risk warrants additional mitigation
2. Document the acceptance decision and supporting analysis
3. Define monitoring requirements for continuous assessment
4. Identify triggers that would necessitate reassessment

Not proceeding where systemic risk is deemed unacceptable

When systemic risk is deemed unacceptable, we will take appropriate steps to bring risk to an acceptable level. This may include implementing additional mitigations until risk is acceptable or halting development while we determine how to proceed safely.

After implementing steps to address unacceptable risk, we will conduct another round of systemic risk analysis and acceptance determination before proceeding.

Staging model development and making available on the market when proceeding

When proceeding with development, market release, or use, we will consider whether a staged approach would help maintain risk below unacceptable levels:

- Limiting initial API access to vetted users
- Gradually expanding access based on post-market monitoring results
- Starting with closed release before any open release
- Using logging systems to track use and safety concerns
- Establishing clear criteria for progressing through stages
- Maintaining the ability to restrict access when necessary

Serious incident response readiness

[List corrective measures to address serious incidents, through improved technical safety and/or security mitigations]

Processes for discovering and assessing risks

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

Planning development

Within four weeks of planning the development or training of a frontier AI model, we will have updated this Framework and begun to assess and mitigate systemic risks.

During development

Development milestones for assessment

We will assess and mitigate systemic risks at the following defined milestones during frontier AI model development:

[select one or two of the below—we recommend 1. and 2f.]

1. **Training compute milestones**: Assessment will occur at each two-fold increase in effective compute to identify potential capability jumps.
2. **Development process milestones**:
    a. After completion of pre-training
    b. Prior to and after each fine-tuning phase
    c. Before reinforcement learning from human feedback
    d. Before expanding model access to new internal teams
    e. Before granting the model additional affordances (network, tool use, etc.)
    f. Before deploying the model to run significant internal operations
3. **Performance-based milestones**:
    a. When evaluation metrics exceed pre-defined thresholds
    b. When the model demonstrates new capabilities on benchmarks

Identification of substantial changes

We implement the following procedures to identify substantial changes in systemic risks:

1. Automated benchmark testing after each significant training phase, and alerting system for unexpected performance increases
2. More thorough evaluations at the milestones identified above

Mitigation preparation

At each milestone, we will assess whether it is reasonably foreseeable that the model will reach higher systemic risk tiers before the next assessment. If such advancement is anticipated, we will:

1. Develop and deploy appropriate mitigations before the model reaches these tiers
2. Document all assessment results and planned mitigations

3. If necessary, pause development to conduct additional evaluation

Documentation standards

For each milestone assessment, we will document:

- Date and development stage
- Capabilities assessment results
- Identified systemic risks
- Mitigation measures implemented or planned
- Decision-making process and conclusions

When significant capability increases are detected that might lead to crossing risk tiers, this documentation will be reviewed by senior leadership before development proceeds.

Safely derived models

Sometimes, we will develop a model by making minor modifications to a model for which we have already conducted the risk assessment outlined in this document. If the reasons for thinking that the original model posed acceptable risk clearly apply to the derivative model, then we may omit some or all of the processes in this document. An example of this would be 'safely derived models' as defined in the General-Purpose AI Code of Practice.

Standards for model evaluations

Selecting or building suitable evaluations

We employ state-of-the-art model evaluations to assess systemic risks and understand the capabilities, propensities, and effects of our frontier AI model. Our selection of methodologies is tailored to the specific systemic risks being evaluated and will include Q&A sets, task-based evaluations, human uplift studies, adversarial testing and model organisms as appropriate.

We maintain evaluation rigor while maximizing efficiency by:

1. Starting with automated screening evaluations
2. Escalating to more intensive evaluations when screening shows potential concerns
3. Reusing evaluation components where appropriate
4. Prioritizing evaluations based on systemic risk tier proximity
5. Employing continuous monitoring for capability shifts

For exploratory research purposes, we may conduct model evaluations with a lower level of rigor, provided that:

1. These evaluations are clearly marked as preliminary
2. Results are not used as the sole basis for determining acceptable risk levels
3. Findings inform more rigorous follow-up evaluations
4. Limitations are transparently documented in our Model Report

External validity

We ensure that our evaluations are representative of the risks they seek to study by:

- Integrating domain experts in the design process
- Implementing appropriate capability elicitation methods (see later section)
- Noting divergence from real-world contexts
- Ensuring diversity and realism in evaluation environments

Internal validity

We maintain high internal validity in our evaluations through:

- Measuring statistical power
- Controlling for confounding variables
- Preventing train-test contamination
- Respecting canary strings

Conducting rigorous evaluations

Scientific and technical standards

We ensure all model evaluations conducted on our frontier AI model maintain high scientific and technical rigor. Our evaluations adhere to quality standards equivalent or superior to those used in scientific peer review in machine learning, natural sciences, or social sciences, as appropriate to the evaluation type.

We aim for the level of rigor demanded by the quality standards for submissions accepted to major machine learning conferences or scientific journals with impact factors in the top quartile of the field.

Elicitation

We ensure all model evaluations of our frontier AI model employ state-of-the-art model elicitation techniques appropriate and proportionate to the systemic risk being assessed. Our elicitation approaches are designed to:

1. Elicit the upper limit of current and reasonably foreseeable capabilities, propensities, and effects
2. Minimize the risk of under-elicitation that could lead to false confidence
3. Identify and prevent model deception during evaluation
4. Match the realistic elicitation capabilities of potential misuse actors in relevant risk scenarios

Our elicitation techniques include, as appropriate for different risk types:

- **Fine-tuning and prompt engineering** to elicit potentially hazardous capabilities or propensities
- **Model scratchpads** for extended reasoning and planning
- **Model ensembles** to amplify capabilities or overcome limitations
- **Scaffolding and tool use** to extend model functionality
- **Grey-box and white-box access** where necessary for thorough evaluation
- **Modified model versions** including base models or helpful-only variants
- **Compute optimization** using appropriate training and inference resources

When elicitation methods increase a model's risk profile during evaluation, we implement enhanced security measures to prevent unauthorized access or harmful model actions.

We calibrate the required amount of elicitation effort based on:

1. The specific systemic risk being assessed
2. Proximity to defined systemic risk tiers
3. The capability profiles of potential threat actors
4. The limitations of safety mitigations being tested

For each evaluation, we document our elicitation approach, justification for the chosen methods, and any limitations encountered during the elicitation process.

Reproducibility

We maintain reproducibility in (at least some of) our evaluations through:

- Enabling successful peer review and reproduction
- Sharing appropriate evaluation data with partners
- Documenting model evaluation code and methodology
- Providing comprehensive documentation of evaluation environments
- Using publicly available APIs and tools where appropriate
- Documenting all elicitation methods

Safety margins

Safety margin implementation

We will implement sufficiently wide safety margins that are:

1. Appropriate to the systemic risks stemming from our models
2. Representative of potential uncertainties, including:
    - Potential under-elicitation during evaluation

    - General uncertainty in assessment and mitigation results
    - Historical inaccuracy of similar risk assessments
3. Accounting for potential model improvements after market release
4. Aligned with state-of-the-art methods and practices

Quantitative safety margins

For model evaluation methods that can be appropriately calibrated with low uncertainty metrics, we will use quantitative safety margins:

- Expressing margins in relative, percentage, or proportional terms
- Calibrating margins based on historical evaluation accuracy
- Adjusting margin size proportionally to the severity of the risk being assessed

Qualitative safety margins

Where quantitative margins are not feasible, we will implement:

- Qualitative safety margins reflecting expert judgment
- Conservative milestone setting in development and deployment plans
- Investments in robust elicitation efforts proportionate to:
    - Potential or actual adversaries in misuse cases
    - Reasonably foreseeable model improvements

Margin methodology

Our approach to safety margins will:

1. Be clearly documented with justifications for the chosen approach
2. Be regularly reviewed and updated based on new information
3. Include different margins for different risk categories when appropriate
4. Incorporate feedback from safety testing and post-deployment monitoring
5. Consider both immediate capabilities and potential growth trajectories

Uncertainty handling

We recognize that uncertainty in risk assessment requires robust margins. Our safety margins will specifically account for:

- Known limitations in our measurement techniques
- Historical accuracy of our capability predictions
- Gaps in current understanding of emerging capabilities
- Potential variation in real-world performance

- Potential failure modes of mitigation strategies

Reassessment for material changes

We will conduct partial or complete reassessment of systemic risk when:

1. There are material changes in how AI systems integrate our frontier AI model
2. New usage patterns emerge that were not accounted for in prior assessments
3. System architecture modifications could affect model behavior or capabilities
4. Inference environment changes could impact safety or security measures

All system integration considerations are documented in our Model Report with explanations of how they influenced our model evaluations and risk assessments.

Representative model evaluations

Our model evaluations will match likely use contexts of our frontier AI model, such as

1. Giving the model access to the tools it is likely to have access to in real situations
2. Using inputs and patterns representative of real situations

Likewise, we will evaluate our models as part of the kinds of systems they are likely to be part of in real deployments.

Stakeholder collaboration

To facilitate adequate consideration of expected use contexts and access relevant expertise, we:

1. Collaborate with relevant actors along the AI value chain
2. Engage with and appropriately remunerate expert and lay representatives from:
    - Civil society organizations
    - Academic institutions
    - Affected communities
    - Industry partners

When stakeholders directly affected by our frontier AI model are not available, we identify and consult suitable representatives to understand their interests and concerns.

For complex or high-impact systemic risks, we establish ongoing advisory relationships with subject matter experts who can guide our evaluation design and interpretation.

Independent external assessors

Assessments before market placement

Starting November 2, 2025, all AI models potentially posing systemic risk (notably excluding safely derived models or those similar to other safe models) will undergo assessment by independent external assessors before market placement.

We will use best efforts to identify and select qualified assessors who:

1. Have significant domain expertise and technical skill in model evaluation
2. Maintain appropriate information security protocols

We will provide assessors with:

1. Sufficient model access as detailed in our evaluation standards
2. Necessary information for effective assessment
3. Adequate time and resources to conduct thorough evaluations

If we cannot identify qualified assessors, we will account for this additional uncertainty in our risk assessment.

Assessments after market placement

After making models available, we will facilitate exploratory independent assessment by:

1. Implementing a research program providing API access to our models in their market form and to models without certain mitigations
2. Allocating free research API credits for systemic risk research
3. Publishing clear criteria for evaluating applications from assessors
4. Contributing to a legal and technical safe harbor for assessors who:
    - Do not intentionally disrupt system availability
    - Do not access sensitive data without consent
    - Do not use findings to threaten stakeholders
    - Adhere to our responsible vulnerability disclosure process

Fully open-sourcing a model provides an alternative means of fulfilling this requirement.

Qualified model evaluation teams and adequate evaluation access and resources

Team composition and qualifications

Our model evaluation teams will be multidisciplinary, combining technical expertise with relevant domain knowledge to ensure comprehensive assessment of systemic risks. Each evaluation team will include members with at least one of the following qualifications, drawn from the General-Purpose AI Code of Practice:

- Research or engineering experience with demonstrable expertise in model evaluation, evidenced by relevant PhDs, peer-reviewed 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 fields directly relevant to the systemic risks being assessed

Model access

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

- Fine-tuning access when needed to properly elicit model capabilities
- Logit access when necessary for thorough evaluation
- Sufficiently high rate limits to enable comprehensive testing
- Access to the model's inputs and outputs, including chain-of-thought reasoning
- Access to models without safeguards when required for proper evaluation
- Additional model affordances such as script execution or browser operation capabilities when needed
- Access to model specifications
- Information about training data
- Results from previous assessments
- Other information relevant to their evaluation tasks

For cases requiring deeper inspection, we will implement:

- Grey-box access offering limited transparency into the model's inner workings
- White-box access providing full visibility of weights, activations, and technical details when necessary

Resource allocation

We will provide evaluation teams with:

1. **Compute resources**:
    - Adequate compute budgets supporting long evaluation runs
    - Capacity for parallel execution and necessary re-runs
    - Scale appropriate to the evaluation requirements
2. **Personnel resources**:
    - Appropriately sized teams for each evaluation task

    - Engineering support for technical implementation
3. **Time resources**:
    - Sufficient time to design, debug, execute, and analyze evaluations rigorously
    - Time allocation proportionate to:
        - The magnitude of the systemic risk being assessed
        - The complexity of the evaluation method
        - The model's stability and performance characteristics
4. **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

Transparency into external input in decision-making

We will consult external actors[12] at various points in making decisions about the development, deployment on the market, or usage of our AI models, including:

- **Pre-scaleup**: Prior to scaling up a frontier AI model, we will provide a pre-scaleup safety case to [external actor]. The pre-scaleup safety case will provide our evidence, based on model evaluations and scaling laws, for why we think that a certain amount of scaling (measured in compute, calendar time, or other quantities) will not cross into risk thresholds for which we do not yet have adequate security mitigations. [External actor] will have … days to review the pre-scaleup safety case and provide nonbinding feedback. We will ensure that before any scaleup, we will have had a pre-scaleup safety case that demonstrates that further scaling is safe.
- **Prior to market deployment**: Before deploying a frontier AI model on the market, we will provide a pre-deployment safety case to [external actor]. The pre-deployment safety case will provide evidence for why our model is safe to deploy externally and will not pose unacceptable systemic risk. [External actor] will have … days to review the pre-deployment safety case and provide nonbinding feedback.
- **Usage**: Before internally deploying a frontier AI model for broad research usage, we will provide an internal usage safety case for review by [external actor].

Alternative: At this time, we are not committing to receiving input from external actors, including government actors, in making decisions about the development, deployment on the market, or usage of our AI models.

Our forecasting

[12] Relevant external actors to consider could include members of the International Network of AI Safety Institutes (Australia, Canada, China, the European Union, France, Japan, Kenya, the Republic of Korea, Singapore, the United Kingdom, and the United States, which have varying levels of government affiliation), industry collectives such as the Frontier Model Forum, and civil society organizations such as Apollo Research and METR.

Current forecasts

Relevant papers:

- 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

Cyber offence

Timeline estimate: We estimate a …% probability of developing models with offensive cyber capabilities sufficient to pose the unacceptable tier within … months, with the following distribution:

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

Justification: …

Underlying assumptions: …

Uncertainty factors: …

CBRN risk

Timeline estimate: We estimate with …% confidence that models will have capabilities sufficient to pose the unacceptable tier for CBRN will emerge within … months, with the following distribution:

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

Justification: …

Underlying assumptions: …

Uncertainty factors: …

Loss of control

Timeline Estimate: We assess with …% confidence that models will have capabilities sufficient to pose the unacceptable tier within … months, with the following distribution:

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

- 90th percentile: …

Underlying assumptions: …

Justification: …

Uncertainty Factors: …

Harmful manipulation

Timeline estimate: We forecast with …% confidence that models will have capabilities sufficient to pose the unacceptable tier within … months, with the following distribution:

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

Underlying assumptions: …

Justification: …

Uncertainty factors: …

Future work

We will aim to integrate model evaluations for our highest systemic risk tiers with forecasting efforts, including:

Scaling law experiments

We will conduct systematic scaling experiments to understand capability trends across:

- Model size parameters
- Training compute allocation
- Inference compute requirements

These experiments will document observed relationships between resource scaling and capability emergence, supporting more accurate prediction of future capabilities and risk.

Model generation comparisons

We will implement structured comparisons across consecutive model generations to:

- Track capability growth trajectories
- Identify acceleration or deceleration in capability development
- Document emergent capabilities not present in previous generations

Additional forecasting techniques

When appropriate, we will employ other techniques to estimate when capabilities, propensities, affordances, and effects might emerge in future models, such as:

- Trend analysis across the wider AI field
- Statistical extrapolation from observed capability patterns
- Structured expert elicitation for capability forecasting
- Systematic comparison with third-party benchmarks and evaluations

All forecasting techniques will be implemented in accordance with relevant state-of-the-art methods and documented to support ongoing refinement of our predictive approaches.

Exploratory research and open-ended red-teaming

Exploratory model evaluation

Recognizing the evolving nature of systemic risk assessment, we conduct exploratory research that goes beyond evaluating known capabilities and risks. Areas we will explore include:

1. Exploratory testing for unanticipated capabilities or emergent behaviors
2. Developing novel evaluation methodologies
3. Investigating potential new systemic risk pathways
4. Researching improved techniques for model elicitation
5. Conducting meta-research on the effectiveness of evaluation methods
6. Supporting forecasting research to anticipate future capabilities

Open-ended red-teaming

We engage in structured, adversarial testing through open-ended red-teaming that:

1. Involves diverse expert and lay representatives from:
    - Civil society
    - Academia
    - Security research
    - Affected communities
2. Employs methodologies designed to discover:

    - Novel attack vectors
    - Unanticipated failure modes
    - Emergent behaviors
    - Capability boundaries
3. Provides appropriate freedom for creative exploration of:
    - Potential misuse scenarios
    - System vulnerabilities
    - Safety limitation circumvention
    - New risk vectors

When stakeholders directly affected by our frontier AI model are not available for red-teaming, we identify and engage suitable representatives to assess potential impacts from their perspective.

Integration of findings

Results from exploratory research and open-ended red-teaming are:

1. Documented and analyzed for systemic risk implications
2. Incorporated into our systemic risk assessments
3. Used to improve future evaluation methodologies
4. Shared with relevant authorities and other stakeholders as appropriate
5. Applied to enhance our risk mitigation strategies

Sharing tools & best practices

Our organization recognizes that systemic risks are influenced by the interactions between multiple AI models and systems, including potential chain reactions of incidents or malfunctions. Effective assessment and mitigation of systemic risks requires collaboration within the broader AI ecosystem.

Our sharing approach

We will share state-of-the-art model evaluation tools and best practices with relevant actors in the AI ecosystem, particularly those engaged in risk assessment or mitigation of AI models or systems that may interact with our frontier AI models. These include:

- Other model providers (especially SMEs)
- Downstream providers
- Independent external assessors
- Academic institutions

Our sharing program will include established methodologies and procedures while maintaining our ability to perform rigorous risk assessment and mitigation for our own models. We will dedicate engineering and support

resources to facilitate this sharing without compromising the effectiveness of our internal systemic risk assessment team.

Sharing mechanisms

Until there are established industry or government processes, we will implement the following mechanisms:

1. **Open sharing**: Where possible, we will publicly release source code, documentation, and training materials to maximize accessibility for all relevant actors.
2. **Secure sharing**: For more sensitive tools or methodologies that require restricted access, we will establish secure channels, such as through industry organizations, to facilitate sharing with qualified recipients.

Continuous improvement

We will regularly review and enhance our sharing program based on:

- Feedback from recipients
- Advances in evaluation methods
- Evolving best practices
- Guidance from relevant authorities

This approach allows us to contribute meaningfully to the broader AI safety ecosystem while maintaining necessary safeguards around sensitive information.

Post-market monitoring

We will conduct post-market monitoring to gather information about model capabilities, propensities, affordances, and effects for assessing and mitigating systemic risk. Our monitoring will:

1. Ensure our models do not pose unacceptable systemic risk as defined by our risk acceptance criteria
2. Support forecasting work

Monitoring methods

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

[list as appropriate, perhaps including some of the below]

- **User feedback collection** through structured channels and surveys
- **Reporting channels** for users to flag potential issues
- **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
- **Academic collaboration** with researchers studying our models' effects
- **Stakeholder dialogues** with affected groups
- **Technical monitoring** where feasible, including:
    - Development of privacy-preserving monitoring techniques
    - Implementation of watermarks and fingerprinting
    - Metadata analysis where technically and legally appropriate

Focus areas

We will specifically collect information about:

1. Breaches of use restrictions and subsequent incidents
2. For closed-source models, aspects relevant to risk assessment that are not transparent to third parties

Internal monitoring

For AI systems incorporating our models that we provide or deploy ourselves, we will monitor the model as part of these systems to effectively assess and mitigate systemic risks.

Third-party information

When our chosen monitoring methods don't provide sufficient information, we will seek cooperation with licensees, downstream providers, and end-users to receive relevant information, always in accordance with applicable data protection laws. For consumer end-users, we will implement opt-in sharing mechanisms for relevant monitoring information.

Documentation

Safety and Security Model Reports

Our Model Reports will detail the risk assessment and mitigations we have conducted for each frontier AI model potentially posing systemic risk.

Level of detail

Our Model Reports will provide a level of detail that:

1. Is proportionate to the level of systemic risk that our models reach or are reasonably foreseen to reach along their lifecycle
2. Allows clear understanding of how we implement systemic risk assessment and mitigation measures

Each Model Report will contain sufficient reasoning to justify these determinations, including explanations when a lower level of detail is appropriate (such as for safely derived models).

Documentation of release decisions

Each Model Report will provide clear reasoning and information justifying our decision to release a model, including:

1. A comprehensive chain of reasoning demonstrating that systemic risk is acceptable according to our acceptance criteria, including explanation of the safety margin incorporated
2. Clear documentation of conditions under which our reasoning would no longer hold
3. Information about independent external assessors' involvement in the decision-making process, including how their evaluations were incorporated

Documentation of systemic risk assessment and mitigation

Our Model Reports will document in detail:

1. Our systemic risk selection process, including the rationale for selecting specific risks
2. The systemic risk estimation process, justifying all decisions made during assessment, particularly regarding qualitative vs. quantitative approaches
3. For safely derived models, comprehensive justification of how criteria for safe originator models and safely derived models are fulfilled
4. Limitations and uncertainties in our risk assessment, including:
    - Safety margins implemented
    - Any lower levels of rigor in evaluation with justification
    - Consequences for model evaluation
5. Model elicitation efforts for each evaluation
6. How AI systems information was incorporated for each evaluation
7. Qualifications of internal teams and external assessors, including access levels and resources provided
8. Comparison of risk levels both with and without mitigations
9. All implemented mitigations and their limitations

Additionally, we will document:

10. The basis for concluding our evaluations are state-of-the-art
11. Internal validity, external validity, and reproducibility of each evaluation
12. Coverage of expected use contexts and modalities
13. Results from exploratory research and open-ended red-teaming
14. Tools, best practices, and data shared with others for model evaluations

External reports

Our Model Reports will include:

1. Available reports from independent external assessors who reviewed our models before market placement and from security reviews
2. Where no external assessor was involved, a clear explanation why criteria necessitating such involvement were not met, or why no suitable assessor was found
3. Where external assessors were involved, justification for their selection based on qualification criteria, unless they have been recognized by regulatory authorities

Algorithmic improvements

Each Model Report will contain high-level information about algorithmic or other improvements specific to the model compared to other models available on the market, where that information is relevant to understanding significant changes in the systemic risk landscape.

Specification of intended model behaviour

Our Model Reports will clearly specify how we intend the model to operate, including:

1. Principles the model is intended to follow
2. How the model prioritizes different types of instructions
3. Topics on which the model is intended to refuse instructions

Model Report updates

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:

1. Significant changes in the model's capabilities through post-training, elicitation, or affordance changes
2. Data drift, improvements in evaluation methods, or factors suggesting previous assessments are no longer accurate
3. Improvements in mitigations
4. Serious incidents
5. Information from post-market monitoring indicating changes in use patterns
6. Information from internal use of the model

Each updated Model Report will:

1. Include content from the previous report, updated where relevant
2. Include new reports from independent external assessors
3. Assess whether our Framework was properly adhered to
4. Include a detailed changelog with version number and date
5. Explain why, if applicable, we conclude there has been no material change in the systemic risk landscape

Adequacy assessments

Adequacy assessment process

When conducting adequacy assessments, we will consider:

1. Best practices across the industry
2. Relevant research and state-of-the-art science
3. Incidents or malfunctions and serious incident reports
4. Available independent external expertise

Completed adequacy assessments will be provided to our management body or another appropriate independent body within 5 business days of completion.

Model-specific adequacy assessment

For each model meeting classification conditions (except safely derived models), we will document:

1. General description of techniques and assets used in development
2. Outcomes of systemic risk identification
3. Available forecasts showing capabilities, systemic risk indicators, and risk tiers the model is expected to reach
4. Available systemic risk analysis results
5. Overview of planned systemic risk analysis, acceptance determination, and technical mitigations
6. Characterization of systemic risks that may arise during development, with assessment of significant risks during development phase

We will update these assessments when reaching defined milestones if:

1. Significant systemic risks arise during development
2. There have been material changes to the above information
3. No assessment has been completed in the past 12 weeks
4. The model has not been made available on the market

Updates will focus on development-phase risks, particularly AI research automation and security safeguards.

Framework adequacy assessment

We will conduct Framework adequacy assessments:

1. When made aware of material changes that could undermine Framework adequacy
2. Every 12 months starting from our first model market placement

These assessments will evaluate:

1. Whether the Framework addresses all required components and is consistent with regulatory requirements
2. Whether the Framework remains proportionate to risks from models we plan to release within the next 12 months
3. Whether there is strong reason to believe the Framework will be adhered to

Retention

We will maintain comprehensive documentation including:

1. Documentation required under other regulatory frameworks
2. Our Framework, including previous versions
3. Model Reports, including previous versions
4. All completed adequacy assessments

All documentation will be retained for at least 12 months after model retirement.

Systemic risk responsibility allocation

Defining responsibilities

We will clearly define responsibilities for managing systemic risk across all organizational levels:

1. **Risk oversight**: The [specific committee] of our management body will oversee systemic risk management activities
2. **Risk ownership**: [Specific management roles] will take direct responsibility for managing systemic risks
3. **Support and monitoring**: [Specific role] will support and monitor systemic risk management, supported by a central risk function
4. **Assurance**: [Specific role] will provide assurance about adequacy of systemic risk management, supported by an independent audit function

Allocation of appropriate resources

Our management will oversee allocation of appropriate resources proportionate to systemic risk levels:

1. Human resources with relevant expertise
2. Financial resources
3. Access to necessary information and knowledge
4. Computational resources for thorough evaluation

Promotion of a healthy risk culture

We will promote a balanced approach to systemic risk, encouraging appropriate risk awareness without excessive risk-seeking or risk-aversion:

1. Setting tone from leadership regarding balanced risk management
2. Enabling effective communication and challenge of risk decisions
3. Creating incentives discouraging excessive risk-taking
4. Regularly surveying staff about risk awareness and comfort raising concerns
5. Maintaining active reporting channels with appropriate follow-up

[add company specifics as relevant]

Serious incident reporting

Methods for serious incident identification

We will implement methods appropriate to our business model to track information about serious incidents:

1. Tracing model outputs using watermarks, metadata, or other provenance techniques
2. Conducting privacy-preserving logging and metadata analysis where feasible
3. Reviewing external information sources including police and media reports, social media, and incident databases
4. Facilitating reporting by downstream providers, users, and third parties

Relevant information for serious incident tracking, documentation, and reporting

We will track, document, and report at least:

1. Start and end dates of incidents
2. Resulting harm and affected groups
3. Chain of events leading to the incident
4. Model version involved

5. Description of material showing our model's involvement
6. Our response actions and plans
7. Recommendations for regulatory response
8. Root cause analysis including outputs, inputs, and safeguard failures
9. Known near-misses

We will investigate causes and effects of incidents to inform future risk analysis, with level of detail proportionate to incident severity.

Reporting timelines

For serious incidents, where agreed with relevant authorities, we will submit initial reports containing information points 1–7 above as soon as possible, and at most:

1. Within 2 days for serious disruption of critical infrastructure
2. Within 10 days for grave harm to physical integrity or suspected causal relationship
3. Within 15 days for serious harm to health, fundamental rights infringements, or serious property/environmental damage
4. Within 5 days for serious cybersecurity incidents or model weight exfiltration

We will submit intermediate reports every 4 weeks until resolution, and final reports within 60 days of resolution, covering all information requirements.

Retention period

We will retain all documentation produced under this commitment for at least 36 months from documentation date or incident date, whichever is later.

Non-retaliation protections

We will not retaliate against any worker providing information about systemic risks to regulatory authorities. We will annually inform workers of designated regulatory channels for receiving such information.

Notifications

We will share information about models potentially posing systemic risks with authorities as required by applicable rules, such as the General Purpose AI Code of Practice, *Notifications* section.

Public transparency

We will publish information about systemic risks that improves:

1. Public awareness and knowledge of risks
2. Societal resilience against risks
3. Detection of risks before and when they materialize

Specifically, we will publish on our website:

1. New or updated versions of our Framework
2. Model Reports with redactions as necessary for information which might increase risk, including sufficient information for public understanding of:
    - How the assessed risk of the model relates to our risk tiers.
    - For risk tiers not reached, an explanation of why the evidence collected shows that the risk tier is not reached.
    - The quality of elicitation.
    - Any reports by third parties.

Published information will include redactions necessary to prevent increased risk from undermining safety or security mitigations or revealing sensitive commercial information disproportionate to the societal benefit.

Improving and updating the Framework

Over time, we will update and improve our Framework with the intent of making it more effective for evaluating and mitigating systemic risks by:

1. Incorporating insights gained from applying the Framework to our frontier AI models
2. Refining and validating forecasts over time by comparing them to historical trends and other estimates
3. Regularly updating the Framework to incorporate state-of-the-art measures and procedures for risk assessment and mitigation

An updated version of the Framework will be considered complete when:

1. It has been reviewed by our safety and AI risk team
2. It has been approved by our management body acting in its supervisory function
3. All changes have been documented with justifications
4. The changelog has been updated with version number and date of change

Systemic risk selection

When planning the development of a new model with the potential to pose systemic risk, we will identify potential systemic risks specific to our frontier AI model's high-impact capabilities by:

1. Reviewing a range of potential risks, such as those outlined in Appendices 1.1 and 1.2 of the General-Purpose AI Code of Practice
2. Analyzing information on risks exhibited by similar models on the market

3. Consulting current research and expert opinions on emerging risks
4. Engaging with stakeholders likely to be affected by our frontier AI model

Determining systemic risk scenarios

For each identified systemic risk, we will develop detailed risk scenarios that:

1. Characterize the type, nature, and sources of the risk
2. Outline pathways to harm, including reasonably foreseeable negligent and intentional misuse
3. Document how the development, making available on the market, and/or use of our frontier AI model could produce the systemic risk
4. Identify the model capabilities, propensities, affordances and contextual factors which would determine whether the risk is present
5. Identify which mitigations would most effectively lead to acceptable risk

Our risk modeling methodology incorporates both quantitative and qualitative approaches, combining threat modeling with specific model evaluations to create comprehensive risk scenarios that inform our subsequent analysis and mitigation strategies.

Changelog

Version	Date	Changes	Justification
1.0	2025-04-07	Initial framework	…