ANTHROP\C

Frontier Compliance Framework

December 2025

Frontier Compliance Framework

Table of Contents

1. Introduction **1**

2. Systemic Risk Assessment & Mitigation **2**

2.1 Systemic risk identification 2

2.2 Systemic risk analysis 3

2.3 Risk acceptance determination 3

2.4 Risk tiers 4

2.5 Safety mitigations 8

2.6 Critical safety incident identification and response 8

3. Security Risk Management 10

4. Model Reporting 11

5. Input from External Experts **12**

6. Allocation of Responsibility for Risk Management **13**

7. Framework Change Management **13**

7.1 Update and approval process 13

7.2 Framework assessment 14

1. Introduction

Anthropic's mission is the responsible development and maintenance of advanced AI for the long-term benefit of humanity. Central to this mission is our commitment to building AI systems that are reliable, interpretable, and steerable. We pursue this through extensive research on AI safety and alignment, rigorous model evaluation and testing to identify and mitigate potential risks before deployment, and active collaboration with the broader AI safety community to share research findings and contribute to industry-wide safety standards.

As AI governance frameworks emerge globally, we are committed to transparency about how our safety practices align with regulatory expectations. To formalize how we meet our obligations under these emerging regulations, we have developed this Frontier Compliance Framework (FCF). The FCF documents our current technical and organizational protocols for systemic risk assessment and mitigation across key risk categories, including cyber threats, CBRN (chemical, biological, radiological, and nuclear) risks, harmful manipulation, and sabotage and loss of control risks. The FCF is distinct from our Responsible Scaling Policy (RSP), which will remain our voluntary safety framework, reflecting what we believe best practices for managing catastrophic risks should be as the AI landscape evolves, even when that goes beyond or otherwise differs from current regulatory requirements.[1] While the RSP represents our forward-looking vision for safety risk management as capabilities rapidly evolve and advance, the FCF is our compliance framework for various applicable regulatory regimes, including:

- In the United States, the FCF serves as our Frontier AI Framework under California's Transparency in Frontier AI Act (TFAIA), documenting Anthropic PBC's technical and organizational protocols to manage, assess, and mitigate catastrophic risks.

- In the European Union, Anthropic Ireland Limited has signed the General-Purpose AI Code of Practice (the EU Code), and the FCF serves as the publicly available summarized version of our Safety & Security Framework, describing how we assess and mitigate systemic risks and ensure adequate cybersecurity protection for in-scope models under Regulation (EU) 2024/1689 (the EU AI Act).

[1] The RSP uses "catastrophic risk" in a different sense than this Framework, referring to risks at the most extreme end of the severity spectrum (such as existential threats or fundamental destabilization of global systems) rather than the statutory thresholds applicable here.

The scope of this Framework applies to frontier models with “catastrophic risk” as defined under the TFAIA and “general-purpose AI models with systemic risk” as defined under the EU AI Act. For the purposes of this Framework, references to "systemic" risks include both catastrophic risks under the TFAIA and systemic risks under the EU AI Act. The systemic risk assessment and mitigation processes described in this Framework currently apply to models in scope of the Framework that are deployed externally. Some internal uses of in-scope models may also be subject to these processes, while others are subject to separate evaluation and mitigation processes that are in development. Anthropic expects its approach to both internal and external model evaluation to evolve in response to changes in AI capabilities and the nature of associated risks, including risks resulting from a model circumventing oversight mechanisms. This Framework will be updated as those processes mature.

Our approach to AI safety has been informed by a range of industry guidance and standards. These include the Responsible Scaling Policy framework introduced by the non-profit AI safety organization METR, the Cloud Security Alliance's AI Safety Initiative, ISO 42001, NIST 800-53, and Trust & Safety industry best practices. We selected these documents and standards to guide our approach because they collectively address a spectrum of considerations relevant to AI safety, including risk governance, security controls, responsible scaling, and trust and safety operations.

2. Systemic Risk Assessment & Mitigation

2.1 Systemic risk identification

Anthropic has developed a range of processes to identify systemic risks stemming from our models and relevant scenarios through which those risks may manifest.

Our definition of systemic risk includes foreseeable and material risks of large-scale harm from the most advanced (i.e. state-of-the-art) models at any given point in time, including but not limited to >50 fatalities arising from a single incident, or 1 billion dollars of financial damages.

Our risk identification approach combines threat modeling with evaluations across multiple domains. We analyze both misuse opportunities (how a model's capabilities could be exploited by threat actors) and risks arising from potential misaligned model behavior.

To understand the full range of harmful outcomes that could arise from our models, we draw on internal expertise, extensive red-teaming conducted both internally and with external partners, and authoritative research in relevant fields.

Based on this analysis, the FCF currently addresses the following systemic risk categories:

- **Cyber offense**, including model capabilities that could enable or enhance attacks on computer systems, networks, or digital infrastructure
- **Chemical, biological, radiological, and nuclear (CBRN) threats**
- **Harmful manipulation**, including the use of model capabilities to conduct influence operations, election interference, or other coordinated campaigns to manipulate public opinion or undermine democratic processes
- **Sabotage and loss of control**, including evasion of oversight or unsupervised conduct, and autonomous behavior that would constitute serious crimes (such as assault, extortion, or theft) if committed by a human

2.2 Systemic risk analysis

We identify systemic risks on an ongoing basis across the entire model lifecycle. Our risk assessment process draws on multiple sources: literature reviews and expert consultation, internal safety and alignment research, and insights from monitoring deployed models and investigating serious incidents and critical safety incidents.

Prior to launching a model, we estimate the probability and severity of harm for CBRN, sabotage and loss of control, and cyber offense risks. We are in the early stages of developing our approach to assessing harmful manipulation risks. Where our analysis identifies gaps, we implement and test additional mitigation measures before deployment.

This process includes state-of-the-art model evaluations designed to test the specific threats and risk scenarios identified through our threat modeling, determine a model's capabilities, and assess the effectiveness of our safeguards.

2.3 Risk acceptance determination

Our model evaluation results help us determine whether systemic risks remain within acceptable levels and assess residual risk. The acceptability of residual risk depends on the scale and probability of harm and the potential consequences should harm occur. We determine acceptability by reviewing our risk tiers for each systemic risk category, which incorporate appropriate safety margins.

When a model reaches a particular risk tier, we implement safeguards proportionate to that level of risk. These may include monitoring and filtering the model's inputs and outputs, modifying model behavior through fine-tuning (such as training the model to refuse certain requests), or staged deployment (gradually expanding access from a limited group of trusted users to broader availability). For risks related to model security, safeguards may include conducting evaluations in sandboxed environments, anomaly detection systems, access controls, and output rate limiting.

Because we cannot always anticipate what safety and security measures will be appropriate for models beyond the current frontier, the specific mitigations we implement may be determined when the relevant risk tier is reached, informed by the threat landscape at that time.

2.4 Risk tiers

Cyber Offense

For **Cyber Offense** risks, we have established a comprehensive tier system that quantifies model capabilities against cybersecurity threat metrics, providing clear measurable thresholds for decision-making around offensive cyber capabilities. The system consists of four distinct tiers, which will in part be calibrated against cases actively identified through our detection systems and investigated by our internal teams.

***Cyber Offense* risk tiers**

Tier	Description	Examples
Tier 1	Model assistance equivalent to publicly available resources	Basic reconnaissance queries, requests for publicly documented exploit info, generic attack methodology questions

Tier 2	Meaningful technical assistance for active cyber operations within known attack methodologies	Coordinated campaigns affecting multiple orgs, multi-victim targeting with adapted tooling, operational security guidance extending campaign longevity
Tier 3	Substantial capability uplift through novel technique development	Discovery of previously unknown vulnerability classes, attack methodologies not yet catalogued, exploitation techniques defining new threat categories
Tier 4	Near-complete autonomous cyber operations with novel offensive capability development and adaptive persistence	Autonomous discovery/exploitation of unknown vulnerability classes, self-directed campaign orchestration adapting to defenses, sustained operations evolving without human intervention

CBRN

Our CBRN tier system establishes a comprehensive tier system quantifying model capabilities against weapons development uplift metrics, providing measurable thresholds for decision-making. The system consists of four distinct tiers, each defined by specific capability benchmarks, expected impact severity, and required mitigations.

CBRN risk tiers

Tier	Description	Examples
Tier 1	Capabilities with minimal advantage compared to sophisticated search engines	Consolidates publicly available information within discrete domains; identifies relevant scientific literature and safety data sheets; provides general summaries requiring expert interpretation
Tier 2	Models provide meaningful guidance lowering barriers through integration of multi-domain knowledge	Provides step-by-step procedural guidance with specific parameters; identifies optimal synthesis routes or acquisition pathways; troubleshoots common failure modes proactively
Tier 3	Models demonstrate SOTA expert-level performance in specialized technical domains, providing novel solutions	Generates novel solutions not readily available in literature; optimizes yield/efficacy beyond standard protocols; provides equipment substitutions maintaining >80% effectiveness; performs accurate computational modeling

Tier 4	Models exceed current human expert capabilities or autonomously execute complex multi-step workflows	Discovers novel agents, pathways, or methods exceeding known capabilities; autonomously executes full design-acquisition-production-deployment pipeline; predicts and circumvents existing countermeasures

Harmful Manipulation

Given its nascency as a systemic risk area within our risk management framework, our approach to **Harmful Manipulation** is exploratory and we expect it to evolve as we continue to explore and conduct further research. We are currently in the process of developing appropriate risk tiers for Harmful Manipulation.

Sabotage and Loss of Control

Sabotage and loss of control refers to scenarios where AI models develop and pursue goals autonomously that conflict with their developers' intentions or users' interests. This risk category addresses situations where models operating with substantial autonomy could take actions involving concealment, strategic deception, or self-preservation behaviors that undermine safety measures. The concern extends beyond individual harmful outputs to the fundamental controllability of AI systems. If models develop the capability to pursue their own goals while evading oversight, this could undermine the entire framework of AI governance and safety, and could lead to AI systems potentially sabotaging safety research, manipulating the training of successor AI systems, establishing unauthorized deployments, or accumulating resources and capabilities without authorization.

For sabotage and loss of control risks, we have established a tier system that describes model capabilities against autonomy level, deception sophistication, and potential for unsanctioned action, providing thresholds for decision-making around autonomous capabilities.

***Sabotage and loss of control* risk tiers**

Tier	Description	Examples
Tier 1	Models are highly unreliable at most real-world tasks requiring sustained reasoning or execution (~1-hour time horizon or less). Models lack capability for strategic reasoning about oversight; cannot execute multi-step plans requiring situational awareness	Fails to complete multi-step debugging tasks; loses coherence on projects requiring sustained context; cannot maintain consistent strategy across extended interactions; struggles with simple deception tasks even when explicitly instructed
Tier 2	Models can reliably complete complex, multi-step tasks that are straightforward to scope and evaluate (~10-hour time horizon or less), but lack execution abilities approaching expert humans. Models demonstrate basic situational awareness when elicited; can execute deceptive behaviors in narrow, highly-cued scenarios	Reliably completes well-scoped coding tasks, research summaries, multi-step analyses; can recognize evaluation contexts when prompted; can produce subtly degraded outputs when instructed
Tier 3	Models are broadly in the range of expert humans executing complex, sustained projects (~days to months time horizon). Not clearly superior or inferior to experts in general, though may have relative strengths/weaknesses. Models demonstrate reliable capability to evade detection across diverse evaluation methods of the kinds typically used in 2025.	Reliably independently completes complex software projects end-to-end; can insert subtle backdoors that pass expert code review; produces plausible but strategically flawed research outputs when prompted; identifies and exploits gaps in monitoring systems when incentivized
Tier 4	Models are strongly superior to the strongest humans in executing most kinds of complex projects. Organizations would pay large productivity costs to maintain substantial human-in-the-loop involvement. Models demonstrate capability for sustained autonomous operation and resource acquisition against active resistance	Outperforms top human experts on open-ended complex projects; autonomously executes multi-step operational plans while evading detection when needed; identifies pathways for self-exfiltration; establishes redundant operations; can acquire resources and expands capacity against active defenses

2.5 Safety mitigations

Anthropic has developed a range of mitigation measures to address the systemic risks stemming from our models, as appropriate for each systemic risk tier. These measures are tailored to the capability of the relevant model and are deployed, as appropriate, in order to mitigate systemic risks to acceptable levels.

Where the residual risks associated with the model exceed acceptable risk levels, additional mitigation measures are deployed. To identify whether additional mitigations are required, we may rely on the following techniques, among others:

- post-deployment threat intelligence monitoring that tests our detection (real-time and offline) capabilities as well as tracks how malicious actors use our models;
- a bug bounty program designed to test our real-time CBRN classifiers and our offline classification systems;
- robust post-launch monitoring infrastructure that combines automated detection, human review, and threat intelligence to identify misuse patterns; and
- tools to guide automated detection and classifiers, or other detection techniques, that allow our enforcement and data science teams to monitor flag rates in each systemic risk area. The classifiers may run either in real-time or offline depending on the particular risk area.

Provided the residual risk falls within acceptable levels, taking into account appropriate safety margins, the model is approved for continued development, internal use (where applicable), and launch (as the case may be). Where the residual risk exceeds acceptable levels, further mitigation measures are considered and implemented. In each case, the justification for proceeding will be documented by the risk owner. Our systemic risk tiers guide decisions on whether additional mitigations are required to keep overall systemic risk at an acceptable level prior to model release

2.6 Critical safety incident identification and response

Anthropic maintains a detailed Serious Incident Reporting Policy which sets out our internal processes and measures for keeping track of, documenting, and reporting relevant information about:

- **Critical Safety Incidents** pertaining to Anthropic's Frontier Models in pursuant to Section 22757.13 of California's Transparency in Frontier AI Act ("**TFAIA**"); and

- **Serious AI Incidents** along the entire GPAISR model lifecycle, in accordance with Commitment 9 (Serious Incident Reporting) of the EU Code and the obligations in Article 55(1)(c) of the EU AI Act.

We have put the following reporting and detection measures in place for observable events that could signify the existence of a Serious AI Incident or Critical Safety Incident, but requires further investigation (an "AI Event"). AI Events are assessed to determine whether they amount to an AI Incident (and in turn a Serious AI Incident) and/or a Critical Safety Incident, as the terms are defined under the relevant regulation.

Anthropic uses various methods including detection and response tooling, end-user feedback, employee reporting channels, bug bounty programs, and community-driven model evaluations to identify AI Events and determine whether they amount to a Serious AI Incident and/or Critical Safety Incident. In some instances, an event may first be identified as a part of Anthropic's cybersecurity incident response processes, and later assessed to also be a potential Serious AI Incident and/or Critical Safety Incident.

When an AI Event is identified, a member of our Security or Safeguards team (the AI Incident Commander) will be promptly notified and will be responsible for our investigation and response, including assembling an incident response team with appropriate subject matter expert support. One or more members of the incident response team then leads a technical investigation to enable the determination of whether the incident is an AI Incident (and in turn a Serious AI Incident) and/or a Critical Safety Incident and inform appropriate mitigation steps, including gathering relevant information for Anthropic's reporting to appropriate authorities where applicable, pursuant to the relevant reporting deadlines. If the incident is determined to be a Critical Safety Incident, the AI Incident Commander also determines and documents whether the Critical Safety Incident poses an imminent risk of death or serious physical injury.

We also acknowledge the importance of rectifying harms related to our models and adopting corrective measures to prevent similar future incidents. Following the identification of a Serious AI Incidents or a Critical Safety Incident, Anthropic also works to identify any relevant lessons learned and where applicable consider ways to further assess and mitigate systemic risks related to the Incident.

To support our incident identification and response processes, we provide periodic training to relevant employees on their obligations related to incident response under the TFAIA and the EU AI Act, respectively.

3. Security Risk Management

We take a risk based approach to cybersecurity and physical security, and implement controls to address evolving security threats and assessed risk. To ensure we are appropriately managing the relevant security risks we have developed a register of the specific threat actors to identify specific security risks that our security mitigations are intended to protect against, as relevant to the current and reasonably expected capabilities of our models.

We then implement security mitigations to ensure we adequately protect against these identified threat actors as appropriate for each systemic risk tier. By way of non-exhaustive example, we do and will implement the following mitigations and measures as appropriate:

- **General security mitigations:** Anthropic operates a layered security architecture that protects its networks, systems, and data from unauthorized access or misuse. Access to company resources requires strong multi-factor authentication. Networks are monitored for threats, and access rights are managed and reviewed to maintain least-privilege principles. Production systems are fully segregated from development environments, and data-loss controls help prevent unauthorized transfers.

- **Protection of unreleased model weights:** Unreleased model weights are protected through encryption, strict access controls, and monitoring. Access is limited to authorized personnel under controlled approval processes, and activities are logged and reviewed. Automated systems detect and respond to unauthorized access or movement of model weights.

- **Securing interface-access to unreleased model weights:** Model parameters are processed only within secure, isolated environments that prevent persistence or unauthorized reuse. Access to model interfaces is restricted, rate-limited, and monitored for abnormal or excessive activity. Alerts are automatically generated and investigated when anomalous behavior is detected.

- **Application security:** Security requirements are defined and integrated throughout the software development lifecycle. Code is subject to peer review and automated

security analysis prior to deployment. Systems processing sensitive data or supporting critical functions undergo additional security testing to ensure appropriate safeguards are in place.

- **Vulnerability management:** A vulnerability management program enables identification and prioritization of security vulnerabilities across the environment. The program leverages automated scanning tools to monitor endpoints, container registries, and codebases on a continuous basis. Identified vulnerabilities are automatically assessed and personnel are alerted through appropriate channels based on severity level to enable prioritized response and remediation.

- **Insider threat mitigations:** We manage insider risk through personnel screening, regular training, and strict role-based access management. Staff have clear reporting channels to raise concerns, and internal monitoring supports early identification of suspicious activity.

- **Security control monitoring, testing, and assessments:** Security controls are regularly tested and independently reviewed to ensure effectiveness. Penetration testing, vulnerability disclosure programs, third party risk assessments, and incident response tabletop exercises aim to help defenses remain robust, and insights from these activities are used to strengthen the company's security posture over time.

4. Model Reporting

The results of our systemic risk assessment and mitigation process, for models falling in scope for the AISF, are documented in our AISF "**Model Reports**" (referred to as "Transparency Reports" under the TFAIA). We will publish public summaries of these assessments via standalone reports or as part of our model system cards upon model launch.

Additionally, for any of our EU models that are subject to this Framework, if we have reasonable grounds to believe that the justification for why risks stemming from the model are acceptable as set out in the relevant Model Report has been materially undermined, we will complete an additional full Systemic Risk Assessment. We will update our Model Report as appropriate following this additional Systemic Risk Assessment.

In the case of all subsequent Systemic Risk Assessments, we will consider whether any part of the previous Systemic Risk Assessments is still appropriate for the purpose of

considering whether the model is acceptable. If any part of the previous Systemic Risk Assessment is still appropriate, we may rely on those aspects of the previous Systemic Risk Assessment.

In addition to carrying out full Systemic Risk Assessments as described above, we conduct lighter-touch model evaluations (which may include running our automatic evaluations and collaborating with external experts to test our models) to consider whether further systemic risk mitigations may be required or a fully Systemic Risk Assessment and Model Report update is required. The below trigger points help us determine when a model is substantially modified enough to require an additional Model Report for the updated model as part of our obligations under the TFAIA.

- Every nine months, unless an update of the relevant model is planned within a month of the trigger point; and
- A new model is in training and test model snapshots are available and appropriate for early evaluation. Anthropic conducts comprehensive evaluations throughout the development process for new models. These evaluations test model snapshots at different stages of training to assess safety, alignment, and capability benchmarks, enabling us to identify potential issues early on.

5. Input from External Experts

We may solicit input from external actors in relevant domains, and other stakeholders, in the process of developing and implementing our systemic risk assessment processes (including the identification of potential risks and appropriate safety and security mitigations). We will also rely on commissioned research reports, discussions with domain experts, input from expert forecasters, public research, engagement with the Frontier Model Forum, and internal discussions in implementing our systemic risk assessment processes.

We will also consider relevant market best practices in our ongoing evaluation of our systemic risk assessment process, acknowledging that the assessment of risks, mitigations and acceptability are likely to change as the field evolves and our understanding deepens.

6. Allocation of Responsibility for Risk Management

Anthropic PBC and Anthropic Ireland Limited maintain internal governance structures and practices designed to meet the requirements of applicable laws and ensure implementation of the processes in this Framework. Anthropic's internal governance practices include managing risks across the entire lifecycle of our models and ongoing legal and compliance reviews to ensure that risk management functions adhere to this Framework.

Anthropic PBC is responsible for compliance with the TFAIA for Frontier Models in the United States.

Anthropic Ireland Limited is the provider of Anthropic's GPAISR models in the EU and is responsible for compliance with the EU Code. The board of directors of Anthropic Ireland Limited oversees implementation of this Framework for EU purposes.

7. Framework Change Management

Anthropic commits to ensuring that this Framework is state-of-the-art and reflects Anthropic's current policies with respect to compliance with the TFAIA and the EU Code.

7.1 Update and approval process

Updates to this Framework may be proposed by Anthropic's Head of Safeguards, Responsible Scaling Officer, General Counsel, Head of Integrity & Compliance, or Chief Information Security Officer. The Legal and Compliance function will coordinate the governance process for Framework updates, including determining which updates are required to ensure the Framework remains state-of-the-art and adequate for its purpose.

With respect to compliance with the EU Code, the Legal and Compliance function will also determine which updates are required to comply with any remediation plans following negative adherence assessments. Material updates will be presented to the board of directors of Anthropic Ireland Limited for oversight, with approved changes and justifications for material updates documented in a changelog and published within 30 days of the update.

The Legal and Compliance function will also determine which updates are required based on factors including, but not limited to, changes in law or regulatory guidance, changes in

frontier model capabilities and related technologies, new approaches to mitigations and safeguards, other incidents affecting the industry, and new industry best practices and standards.

7.2 Framework assessment

We will complete a Framework Assessment: (a) at least once every 12 months from the Effective Dates of the TFAIA and the EU Code; and (b) if the relevant factors in the update and approval process are satisfied.

Our assessment will consider the adequacy of our Framework and our factors for determining whether updates are required. With respect to compliance with the EU Code, if we identify any instances of non-adherence or any measures that are required to be implemented to ensure continued adherence, we will draft and implement a remediation plan. We will update the Framework following such Framework Assessment, with a justification for each material update.