Anthropic's Responsible Scaling Policy

Anticipating and securing against emerging threats
that accompany increasingly powerful models

Last updated May 14, 2025

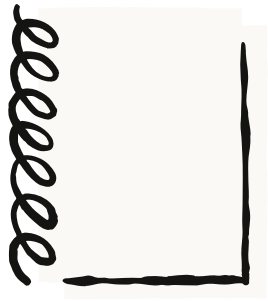
Related:

Read Anthropic's Responsible Scaling Policy →

As frontier AI models advance, we believe they will bring about transformative benefits for our society and economy. AI could accelerate scientific discoveries, revolutionize healthcare, enhance our education system, and create entirely new domains for human creativity and innovation. Frontier AI models also, however, present new challenges and risks that warrant careful study and effective safeguards.

In September 2023, we released the first version of our Responsible Scaling Policy (RSP). We believe that risk governance in this rapidly evolving domain should be proportional, iterative, and exportable. In that spirit, we have continued to refine the RSP over time and will use this page to inform the public of any developments.

Current and Prior Versions



Read the current version of the RSP
Version 2.2 (effective May 14, 2025)

See the PDF

- Version 2.2 and redline (effective May 14, 2025)
- Version 2.1 (effective March 31, 2025)
- Version 2.0 (effective October 15, 2024)
- Version 1.0 (effective September 19, 2023)

May 14, 2025

Version 2.2 implements a minor revision, amending a footnote to exclude both sophisticated insiders and state-compromised insiders from the scope of the ASL-3 Security Standard. Previously, only “highly sophisticated state-compromised insiders” were explicitly excluded. For more details, please see the changelog appended to the policy.

March 31, 2025

Version 2.1 reflected minor updates clarifying which Capability Thresholds would require enhanced safeguards beyond our current ASL-3 standards. First, we have added a new capability threshold related to CBRN development, which defines capabilities that could substantially uplift the development capabilities of moderately resourced state programs. Second, we have disaggregated our existing AI R&D capability thresholds, separating them into two distinct levels (the ability to fully automate entry-level AI research work, and the ability to

cause dramatic acceleration in the rate of effective scaling) and have provided additional detail on the corresponding Required Safeguards. Finally, we have adopted a general commitment to reevaluate our Capability Thresholds whenever we upgrade to a new set of Required Safeguards.

October 15, 2024

Planned ASL-3 Safeguards

In our Responsible Scaling Policy, reaching certain Capability Thresholds requires us to upgrade our safeguards to the ASL-3 Security Standard or the ASL-3 Deployment Standard. Our RSP contains requirements for meeting these standards, and also says we will publicly release key information related to the evaluation and deployment of our models (not including sensitive details). Below, we give non-binding descriptions of our future ASL-3 safeguard plans. We hope sharing these plans will offer useful insights for organizations working on similar systems, and contribute to conversations about emerging best practices.

Deployment Safeguards

Our teams are currently developing and building ASL-3 Deployment Safeguards to mitigate catastrophic risks associated with more advanced models. These safeguards are being designed to prevent misuse of capabilities that could enable severe harm, particularly in relation to chemical, biological, radiological, and nuclear (CBRN) technologies. This overview outlines the planned technical architecture and design of these safeguards, with a focus on the technical aspects and safety considerations of this developing system.

Multi-layered defense-in-depth architecture

Our deployment safeguards will employ a defense-in-depth strategy with four main layers, each designed to catch potential misuse that might pass through previous barriers. The four layers will be:

1. Access controls to tailor safeguards to the deployment context and group of expected users.
2. Real-time prompt and completion classifiers and completion interventions for immediate online filtering.

3. Asynchronous monitoring classifiers for a more detailed analysis of the completions for threats.
4. Post-hoc jailbreak detection with rapid response procedures to quickly address any threats.

Access controls

General access to AI models (e.g. Claude.ai and our API) will use our standard safeguards. However, we recognize that in certain cases, such as safety testing, it may be necessary to tailor these safeguards. To balance security concerns with the need for flexibility, we are developing a tiered access system that allows for nuanced control over safeguard adjustments. At the core of this system will be an enhanced due diligence process under which we will evaluate potential partners based on two key criteria: their overall trustworthiness and the beneficial nature of their use-case. This vetting process will act as a compensating control when a partner's use-case requires adjusting the standard deployment of our safeguards.

Real-time prompt and completion classifiers

Our online prompt and completion classifiers are machine learning models that will analyze user inputs and AI-generated outputs in real-time. The completion classifiers will use a streaming implementation, whereby the classifier updates the score as tokens are generated, rather than waiting for the entire completion. This approach aims to minimize delays to the end user while maintaining the safety standards.

In order to stay up to date on newly discovered harmful patterns, jailbreaks, and obfuscation techniques, the classifiers will be regularly updated using data from other elements of our deployment safeguards including our asynchronous monitoring system and insights from our incident response protocol, bug bounty program, and internal and external red-teaming efforts.

Asynchronous monitoring classifiers

Asynchronous classification will allow us to perform computationally intensive evaluations to provide a deeper level of scrutiny than the real-time classifiers alone, without impacting user-facing latency.

To allow for flexible monitoring processes, our monitoring system will be organized like a flowchart. One implementation could start with simpler AI models like Claude 3 Haiku to quickly and economically scan content and trigger a detailed analysis with an advanced model like Claude 3.5 Sonnet if anything suspicious is found. This setup will be designed to be adapted and updated easily to respond to new threats by adding new AI models or analysis steps.

Post-hoc jailbreak detection

Our jailbreak rapid response procedures will be designed to identify and mitigate attempts to bypass the sets of safety measures described above.

A new jailbreak could be detected by the asynchronous classifiers or through the bug-bounty program. The rapid-response protocol will involve mitigating the impact of the jailbreak by initiating patching. In an instance where a jailbreak patch is unable to be immediately implemented, we will maintain the capability to adjust the model's prompting to reinforce safety constraints. For edge cases or situations requiring human judgment, there will be protocols for escalation to human reviewers.

Our process for minimizing jailbreak risks will include rapid retraining, validation, and testing of classifiers against newly discovered patterns and real-world scenarios. The effectiveness of jailbreak detection and response will be enhanced by collaboration and information sharing. To this end, we are developing a rapid response process for sharing threat intelligence with relevant partners.

Ongoing work

By implementing multiple defense layers, from real-time classifiers to asynchronous monitoring and rapid response systems, we hope to create a deployment safeguards infrastructure that is adaptable to various deployment scenarios.

The success of this system will depend on the integration and improvement of individual components. As implementation of these safeguards progresses, new challenges and opportunities for innovation in AI safety are likely to emerge.

Ongoing research into areas such as balancing security and model utility, improving scalability and performance, enhancing adversarial robustness, and maintaining cross-platform consistency will be crucial in addressing these challenges.

Security Safeguards

The following sections outline the key security controls we plan to implement as part of our ASL-3 safeguards. Many of these security measures are already in place and are dedicated to enhancing our existing ASL-2 controls. In presenting this information, we have carefully balanced the need for transparency with the imperative to protect sensitive security details. As such, this overview provides a high-level description of our planned controls, offering insight into our security approach without compromising our defenses against potential threats.

1. Access management and compartmentalization: Implement multiple clearance levels, data classification, and granular, per-role permission sets to govern employee access to sensitive assets and data including training techniques and hyperparameters. Use multi-tier compartmentalization controls based on asset type to limit blast radius from attacks and reduce insider risk. Educate employees on insider risk and establish a structured insider threat program to incentivize risk reporting.
2. Researcher tooling security: Enhance usability and security of research tools by preventing unnecessary access and limiting user privileges to only what is essential. Provide robust, user-friendly tooling without direct access to model weights.
3. Software inventory management: Create a comprehensive software inventory list through automated scanning. Implement strict software inventory management to track all software components used in development and deployment. Regularly monitor inventory to effectively manage risks and maintain a secure environment.
4. Software supply chain security: Monitor publicly known critical security issues in third-party dependencies. Conduct consistent scanning of all third-party dependencies and maintain a comprehensive vulnerability data repository. Proactively scan all incoming packages to enable download of only secure packages and reduce risk of introducing vulnerabilities.
5. Artifact integrity and provenance: Leverage frameworks (e.g., SLSA) to improve security controls for software build and deployment. Require

packages to have provenance metadata.

6. Binary authorization for endpoints: Enforce binary authorization on every endpoint to allow only trusted, approved software to run. Follow a strict process for approving any software that needs local installation on machines. Use centrally managed execution policies to prevent unauthorized or malicious code. Limit attacker actions through strict security policies and detailed logging for swift incident detection and response.
7. Endpoint patching: Use device management software (e.g., Mobile Device Management) to update all devices to the latest, most secure OS and software versions. Monitor for publicly disclosed vulnerabilities in third-party software and quickly deploy patches or mitigate risks. Restrict access for devices that fall behind on critical updates until patches are applied. Implement automated patching processes to close any manual oversight gaps.
8. Hardware procurement: Source employee endpoints directly from vetted manufacturers for secure chain of custody. Maintain a curated list of approved, recommended hardware and peripherals and provide employees with securely sourced peripherals.
9. Executive Risk Council: Establish an Executive Risk Council sponsored by executive leadership to oversee security programs. Follow a risk-based approach aligned with ISO 27001 standard. Perform periodic reviews to assess our security program's adherence to obligations deriving appropriately from both internal factors (e.g., adopted industry practices, contractual commitments, internal policies) and appropriate external factors (e.g., regulations and statutes).
10. Access control for model weights: Implement multi-party authorization and mandatory code review on production code to remove persistent, high-privilege access to model weights. Grant temporary access and only via the smallest set of necessary permissions to reduce risk of weights exfiltration. Require hardware authentication device prompt, justification and employee approval to grant access.
11. Infrastructure policy: Require all new production infrastructure to be defined in Infrastructure As Code (IaC) before promotion to production environments. Require infrastructure changes be reviewed by the Security team.
12. Cloud security posture management: Reduce risk of compromise due to cloud misconfiguration by defining and implementing internal standards

and best practices. Conduct regular audits of cloud environments and promptly remediate any identified gaps.

13. Red teaming and penetration testing: Partner with a diverse range of external red team and penetration testing experts. Simulate sophisticated attacks, including insider threat and software supply chain compromise scenarios, to identify vulnerabilities. Build an in-house security team with wide-ranging expertise, including APT defense, insider risk, incident response, and secure design.
14. Centralized log management and analysis: Centralize storage of major security-centric log sources in a SIEM/SOAR platform. Enable manual and automated analysis, log retention, rule creation and execution for detections and response workflows. Implement orchestration and response playbooks for automated alert investigation. Use a casebook workflow for security analysts to manage incidents.
15. Access monitoring for critical assets: Conduct manual monitoring of access to model weights and high value IP, including recurring automated detections. Build automated detection across all major log sources for access to critical assets. Leverage additional threat intelligence to continuously improve detections.
16. Deception technology: Install and monitor honeypots (including fake model weights) for high precision detection of unauthorized system access. Implement deception technology in a realistic way to trick attackers and gain insights into their tactics.
17. Physical security: Conduct regular Technical Surveillance Countermeasures (TSCM) at physical spaces using advanced detection equipment and techniques. Tailor TSCM sweeps to specific events, threats or incidents that may trigger an inspection. Regularly sweep physical premises for intruders and conduct physical security red-teaming

Planned Capability Assessments

We plan to publish additional details on our capability assessment methodology in the near future. For information about our past capability assessments, please see our overviews for Claude 3.0 Opus and Claude 3.5 Sonnet.

Learning from Experience

We have learned a lot in our first year with the previous RSP in effect, and are using this update as an opportunity to reflect on what has worked well and what makes sense to update in the policy. As part of this, we reviewed how well we adhered to the framework and identified a small number of instances where we fell short of meeting the full letter of its requirements. These areas were:

1. Our most recent evaluations were completed 3 days later than the 3-month interval. This delay allov **Skip to footer** t ıeir capability elicitation, resulting in higher-quality evaluations. To resolve this issue, the new policy clarifies the ambiguous definition of the evaluation interval, and extends the interval to 6 months to avoid lower-quality, rushed elicitation.
2. In our most recent evaluations, we updated our autonomy evaluation from the specified placeholder tasks, even though an ambiguity in the previous policy could be interpreted as also requiring a policy update. We believe the updated evaluations provided a stronger assessment of the specified "tasks taking an expert 2-8 hours" benchmark. The updated policy resolves the ambiguity, and in the future we intend to proactively clarify policy ambiguities.
3. Some of our evaluations lacked some basic elicitation techniques such as best-of-N or chain-of-thought prompting. Our internal Capability Report discloses these limitations, and we believe the risk of substantial under-elicitation is low. We are now systematically tracking these gaps to avoid future under-elicitation.
4. Our evaluations in one domain were not explicitly designed to establish the 6x scaling buffer mentioned in the previous policy. Since our current techniques aren't capable of giving confident buffer predictions, we instead relied on empirical observations and rough predictions. We believe this presents minimal risk at present, where capability improvements happen more smoothly with scale, but in the future, quantitative predictions may become necessary. We have updated the new policy to note this situation.

In all cases, we found these instances posed minimal risk to the safety of our models. From our review, we learned two valuable lessons to incorporate into our updated framework: we needed to incorporate more flexibility into our policies, and we needed to improve our process for tracking compliance with the RSP.

Since we first released the RSP a year ago, our goal has been to offer an example of a framework that others might draw inspiration from when crafting their own

AI risk governance policies. We hope that proactively sharing our experiences implementing our own policy will help other companies in implementing their own risk management frameworks and contribute to the establishment of best practices across the AI ecosystem.

Education

Financial services

Government

Life sciences

Claude Developer Platform

Overview

Developer docs

Pricing

Amazon Bedrock

Google Cloud's Vertex AI

Console login

Learn

Courses

Connectors

Customer stories

Engineering at Anthropic

Events

Powered by Claude

Service partners

Startups program

Company

Anthropic

Careers

Economic Futures

Research

News

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support center

Terms and policies

Privacy choices

Privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy