ANTHROP\C

Responsible Scaling Policy

Effective October 15, 2024

Supplementary info available at www.anthropic.com/rsp-updates

Executive Summary

In September 2023, we released our Responsible Scaling Policy (RSP), a public commitment not to train or deploy models capable of causing catastrophic harm unless we have implemented safety and security measures that will keep risks below acceptable levels. We are now updating our RSP to account for the lessons we've learned over the last year. This updated policy reflects our view that risk governance in this rapidly evolving domain should be proportional, iterative, and exportable.

Background. AI Safety Level Standards (ASL Standards) are a set of technical and operational measures for safely training and deploying frontier AI models. These currently fall into two categories: Deployment Standards and Security Standards. As model capabilities increase, so will the need for stronger safeguards, which are captured in successively higher ASL Standards. At present, all of our models must meet the ASL-2 Deployment and Security Standards. To determine when a model has become sufficiently advanced such that its deployment and security measures should be strengthened, we use the concepts of Capability Thresholds and Required Safeguards. A Capability Threshold tells us *when* we need to upgrade our protections, and the corresponding Required Safeguards tell us *what standard* should apply.

Capability Thresholds and Required Safeguards. The Required Safeguards for each Capability Threshold are intended to mitigate risk to acceptable levels. This update to our RSP provides specifications for Capabilities Thresholds related to Chemical, Biological, Radiological, and Nuclear (CBRN) weapons and Autonomous AI Research and Development (AI R&D) and identifies the corresponding Required Safeguards.

Capability assessment. We will routinely test models to determine whether their capabilities fall sufficiently far below the Capability Thresholds such that the ASL-2 Standard remains appropriate. We will first conduct preliminary assessments to determine whether a more comprehensive evaluation is needed. For models requiring comprehensive testing, we will assess whether the model is unlikely to reach any relevant Capability Thresholds absent surprising advances in widely accessible post-training enhancements. If, after the comprehensive testing, we determine that the model is sufficiently below the relevant Capability Thresholds, then we will continue to apply the ASL-2 Standard. If, however, we are unable to make the required showing, we will act as though the model has surpassed the Capability Threshold. This means that we will both upgrade to the ASL-3 Required Safeguards and conduct a follow-up capability assessment to confirm that the ASL-4 Standard is not necessary.

Safeguards assessment. To determine whether the measures we have adopted satisfy the ASL-3 Required Safeguards, we will conduct a safeguards assessment. For the ASL-3 Deployment Standard, we will evaluate whether it is robust to persistent attempts to misuse the capability in question. For the ASL-3 Security Standard, we will evaluate whether it is highly protected against non-state attackers attempting to steal model weights. If we determine that we have met the ASL-3 Required Safeguards, then we will proceed to deployment, provided we have also conducted a follow-up capability assessment.

Follow-up capability assessment. In parallel with upgrading a model to the ASL-3 Required Safeguards, we will conduct a follow-up capability assessment to confirm that further safeguards are not necessary. We are currently working on defining any further Capability Thresholds that would mandate ASL-4 Required Safeguards.

Deployment and scaling outcomes. We may deploy or store a model if either of the following criteria are met: (1) the model's capabilities are sufficiently far away from the existing Capability Thresholds, making

the current ASL-2 Standard appropriate; or (2) the model's capabilities have surpassed the existing Capabilities Threshold, but we have implemented the ASL-3 Required Safeguards and conducted the follow-up capability assessment. In any scenario where we determine that a model requires ASL-3 Required Safeguards but we are unable to implement them immediately, we will act promptly to reduce interim risk to acceptable levels until the ASL-3 Required Safeguards are in place.

Governance and transparency. To facilitate the effective implementation of this policy across the company, we commit to several internal governance measures, including maintaining the position of Responsible Scaling Officer, establishing a process through which Anthropic staff may anonymously notify the Responsible Scaling Officer of any potential instances of noncompliance, and developing internal safety procedures for incident scenarios. To advance the public dialogue on the regulation of frontier AI model risks and to enable examination of our actions, we will also publicly release key materials related to the evaluation and deployment of our models with sensitive information removed and solicit input from external experts in relevant domains.

Contents

Introduction

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 our Responsible Scaling Policy (RSP), a first-of-its-kind public commitment not to train or deploy models capable of causing catastrophic harm unless we have implemented safety and security measures that will keep risks below acceptable levels. Our RSP serves several purposes: it is an internal operating procedure for investigating and mitigating these risks and helps inform the public of our plans and commitments. We also hope it will serve as a prototype for other companies looking to adopt similar frameworks and, potentially, inform regulators about possible best practices.

We are now updating our RSP to account for the lessons we've learned over the last year. This policy reflects our view that risk governance in this rapidly evolving domain should be **proportional, iterative, and exportable.**

First, our approach to risk should be proportional. Central to our policy is the concept of AI Safety Level Standards: technical and operational standards for safely training and deploying frontier models that correspond with a particular level of risk. By implementing safeguards that are proportional to the nature and extent of an AI model's risks, we can balance innovation with safety, maintaining rigorous protections without unnecessarily hindering progress. This approach also enables us to allocate resources efficiently, focusing our most stringent safeguards on the models that pose the greater risk, while affording more flexibility for lower-risk systems.

Second, our approach to risk should be iterative. Since the frontier of AI is rapidly evolving, we cannot anticipate what safety and security measures will be appropriate for models far beyond the current frontier. We will thus regularly measure the capability of our models and adjust our safeguards accordingly. Further, we will continue to research potential risks and next-generation mitigation techniques. And, at the highest level of generality, we will look for opportunities to improve and strengthen our overarching risk management framework.

Third, our approach to risk should be exportable. To demonstrate that it is possible to balance innovation with safety, we must put forward our proof of concept: a pragmatic, flexible, and scalable approach to risk governance. By sharing our approach externally, we aim to set a new industry standard that encourages widespread adoption of similar frameworks. In the long term, we hope that our policy may offer relevant insights for regulation. In the meantime, we will continue to share our findings with policymakers.

Although this policy focuses on catastrophic risks, they are not the only risks that we consider important. Our Usage Policy sets forth our standards for the use of our products, including prohibitions on using our models to spread misinformation, incite violence or hateful behavior, or engage in fraudulent or abusive practices, and we continually refine our technical measures for enforcing our trust and safety standards at scale. Further, we conduct research to understand the broader societal impacts of our models. Our Responsible Scaling Policy complements our work in these areas, contributing to our understanding of current and potential risks.

At Anthropic, we are committed to developing AI responsibly and transparently. Since our founding, we have recognized the importance of proactively addressing potential risks as we push the boundaries of AI capability and of clearly communicating about the nature and extent of those risks. We look forward to

continuing to refine our approach to risk governance and to collaborating with stakeholders across the AI ecosystem.

This policy is designed in the spirit of the Responsible Scaling Policy (RSP) framework introduced by the non-profit AI safety organization METR, as well as emerging government policy proposals in the UK, EU, and US. This policy also helps satisfy our Voluntary White House Commitments (2023) and Frontier AI Safety Commitments (2024). We extend our sincere gratitude to the many external groups that provided invaluable guidance on the development and refinement of our Responsible Scaling Policy. We actively welcome feedback on our policy and suggestions for improvement from other entities engaged in frontier AI risk evaluations or safety and security standards. To submit your feedback or suggestions, please contact us at rsp@anthropic.com.

1. Background

AI Safety Level Standards (ASL Standards) are core to our risk mitigation strategy. An ASL Standard is a set of technical and operational measures for safely training and deploying frontier AI models. As model capabilities increase, so will the need for stronger safeguards, which are captured in successively higher ASL Standards. Definitions of ASL Standards and other key terms are available in Appendix A.

The types of measures that compose an ASL Standard currently fall into two categories—Deployment Standards and Security Standards—which map onto the types of risks that frontier AI models may pose.

- **Deployment Standards:** Deployment Standards are technical, operational, and policy measures to ensure the safe usage of AI models by external users (i.e., our users and customers) as well as internal users (i.e., our employees). Deployment Standards aim to strike a balance between enabling beneficial use of AI technologies and mitigating the risks of potentially catastrophic cases of misuse.

- **Security Standards:** Security Standards are technical, operational, and policy measures to protect AI models—particularly their weights and associated systems—from unauthorized access, theft, or compromise by malicious actors. Security Standards are intended to maintain the integrity and controlled use of AI models throughout their lifecycle, from development to deployment.

We expect to continue refining our framework in response to future risks (for example, the risk that an AI system attempts to subvert the goals of its operators).

At present, all of our models must meet the ASL-2 Deployment and Security Standards. The ASL-2 Security and Deployment Standards provide a baseline level of safe deployment and model security for AI models. These standards, which are summarized below, are available in full in Appendix B.

- The ASL-2 Deployment Standard reduces the prevalence of misuse, and includes the publication of model cards and enforcement of Usage Policy; harmlessness training such as Constitutional AI and automated detection mechanisms; and establishing vulnerability reporting channels as well as a bug bounty for universal jailbreaks.

- The ASL-2 Security Standard requires a security system that can likely thwart most opportunistic attackers and includes vendor and supplier security reviews, physical security measures, and the use of secure-by-design principles.

Although the ASL-2 Standard is appropriate for all of our current models, that may not hold true in the

future as our models become more capable. To determine when a model has become sufficiently advanced such that its deployment and security measures should be strengthened, we use the concepts of Capability Thresholds and Required Safeguards.

A Capability Threshold tells us *when* we need to upgrade our protections, and the corresponding Required Safeguards tell us *what standard* should apply. A Capability Threshold is a prespecified level of AI capability that, if reached, signals (1) a meaningful increase in the level of risk if the model remains under the existing set of safeguards (2) a corresponding need to upgrade the safeguards to a higher ASL Standard. In other words, a Capability Threshold serves as a trigger for shifting from an ASL-N Standard to an ASL-N+1 Standard (or, in some cases, moving straight to ASL N+2 or higher). Depending on the Capability Threshold, it may not be necessary to upgrade both the Deployment and Security Standards; each Capability Threshold corresponds to specific Required Safeguards that identify which of the ASL Standards must be met.

2. Capability Thresholds and Required Safeguards

Below, we specify the Capability Thresholds and their corresponding Required Safeguards. The Required Safeguards for each Capability Threshold are intended to mitigate risk from a model with such capabilities to acceptable levels. In developing these standards, we have weighed the risks and benefits of frontier model development. We believe these safeguards are achievable with sufficient investment and advance planning into research and development and would advocate for the industry as a whole to adopt them. We will conduct assessments to inform when to implement the Required Safeguards (see Section 4). The Capability Thresholds summarized below are available in full in Appendix C.

Capability Thresholds	Required Safeguards
Chemical, Biological, Radiological, and Nuclear (CBRN) weapons. The ability to significantly help individuals or groups with basic technical backgrounds (e.g., undergraduate STEM degrees) create/obtain and deploy CBRN weapons.	This capability could greatly increase the number of actors who could cause this sort of damage, and there is no clear reason to expect an offsetting improvement in defensive capabilities. The ASL-3 Deployment Standard and the ASL-3 Security Standard, which protect against misuse and model-weight theft by non-state adversaries, are required.
Autonomous AI Research and Development (AI R&D): The ability to either fully automate the work of an entry-level remote-only Researcher at Anthropic, or cause dramatic acceleration in the rate of effective scaling.	This capability could greatly increase the pace of AI development, potentially leading to rapid and unpredictable advances in AI capabilities and associated risks. At minimum, the ASL-3 Security Standard is required, although we expect a higher security standard (which would protect against model-weight theft by state-level adversaries) will be required, especially in the case of dramatic acceleration. We also expect a strong affirmative case (made with accountability for both the reasoning and implementation) about the risk of models pursuing misaligned goals will be required.[1]

[1] *We will specify these requirements more precisely when we reach the 2-8 hour software engineering tasks checkpoint.*

We will consider it sufficient to rule out the possibility that a model has surpassed the Autonomous AI R&D Capability Threshold by considering an earlier (i.e., less capable) checkpoint: the ability to autonomously perform a wide range of 2-8 hour software engineering tasks. We would view this level of capability as an important checkpoint towards both Autonomous AI R&D as well as other capabilities that may warrant similar attention (for example, autonomous replication). We will test for this checkpoint and, by the time we reach it, we aim to have met (or be close to meeting) the ASL-3 Security Standard as an intermediate goal, and we will share an update on our progress around that time. At that point, we will also specify Required Safeguards for this Capability Threshold in more detail, update our list of Capability Thresholds to consider additional risks that may arise, and test for the full Autonomous AI R&D Capability Threshold and any additional risks.

We will also maintain a list of capabilities that we think require significant investigation and may require stronger safeguards than ASL-2 provides. This group of capabilities could pose serious risks, but the exact Capability Threshold and the Required Safeguards are not clear at present. These capabilities may warrant a higher standard of safeguards, such as the ASL-3 Security or Deployment Standard. However, it is also possible that by the time these capabilities are reached, there will be evidence that such a standard is not necessary (for example, because of the potential use of similar capabilities for defensive purposes). Instead of prespecifying particular thresholds and safeguards today, we will conduct ongoing assessments of the risks with the goal of determining in a future iteration of this policy what the Capability Thresholds and Required Safeguards would be.

At present, we have identified one such capability:

Capabilities	**Ongoing Assessment**
Cyber Operations: The ability to significantly enhance or automate sophisticated destructive cyber attacks, including but not limited to discovering novel zero-day exploit chains, developing complex malware, or orchestrating extensive hard-to-detect network intrusions.	This will involve engaging with experts in cyber operations to assess the potential for frontier models to both enhance and mitigate cyber threats, and considering the implementation of tiered access controls or phased deployments for models with advanced cyber capabilities. We will conduct either pre- or post-deployment testing, including specialized evaluations. We will document any salient results alongside our Capability Reports (see Section 3).[2]

Overall, our decision to prioritize the capabilities in the two tables above is based on commissioned research reports, discussions with domain experts, input from expert forecasters, public research, conversations with other industry actors through the Frontier Model Forum, and internal discussions. As the field evolves and our understanding deepens, we remain committed to refining our approach.[3]

[2] *We hope to publish updates approximately every 6 months.*
[3] *We recognize the potential risks of highly persuasive AI models. While we are actively consulting experts, we believe this capability is not yet sufficiently understood to include in our current commitments.*

3. Capability Assessment

3.1. Preliminary Assessment

We will routinely test models to determine whether their capabilities fall sufficiently far below the Capability Thresholds such that we are confident that the ASL-2 Standard remains appropriate. We will first conduct preliminary assessments (on both new and existing models, as needed) to determine whether a more comprehensive evaluation is needed. The purpose of this preliminary assessment is to identify whether the model is notably more capable than the last model that underwent a comprehensive assessment.

The term "notably more capable" is operationalized as at least one of the following:

1. The model is notably more performant on automated tests in risk-relevant domains (defined as 4x or more in Effective Compute[4]).
2. Six months' worth of finetuning and other capability elicitation methods have accumulated.[5] This is measured in calendar time, since we do not yet have a metric to estimate the impact of these improvements more precisely.[6]

In addition, the Responsible Scaling Officer may in their discretion determine that a comprehensive assessment is warranted.

If a new or existing model is below the "notably more capable" standard, no further testing is necessary.

3.2. Comprehensive Assessment

For models requiring comprehensive testing, we will assess whether the model is unlikely to reach any relevant Capability Thresholds absent surprising advances in widely accessible post-training enhancements.[7] To make the required showing, we will need to satisfy the following criteria:

1. **Threat model mapping:** For each capability threshold, make a compelling case that we have mapped out the most likely and consequential threat models: combinations of actors (if relevant), attack pathways, model capability bottlenecks, and types of harms. We also make a compelling case that there does not exist a threat model that we are not evaluating that represents a substantial amount of risk.

---

[4] *"Effective Compute" is a scaling-trend-based metric that accounts for both FLOPs and algorithmic improvements. An Effective Compute increase of K represents a performance improvement from a pretrained model on relevant task(s) equivalent to scaling up the baseline model's training compute by a factor of K. We plan to track Effective Compute during pretraining on a weighted aggregation of datasets relevant to our Capability Thresholds (e.g., coding and science). This is, however, an open research question, and we will explore different possible methods. More generally, the Effective Compute concept is fairly new, and we may replace it with another metric in a similar spirit in the future.*

[5] *This is a broad category, including techniques like improved prompting and agent scaffolding.*

[6] *Exploring ways to integrate these types of improvements into an overall metric is an ongoing area of research.*

[7] *By "widely accessible," we mean techniques that are available to a moderately resourced group (i.e., do not involve setting up large amounts of custom infrastructure or using confidential information). We include headroom to account for the possibility that the model is either modified via one of our own finetuning products or stolen in the months following testing, and used to create a model that has reached a Capability Threshold. That said, estimating these future effects is very difficult given the state of research today.*

2. **Evaluations:** Design and run empirical tests that provide strong evidence that the model does not have the requisite skills; explain why the tests yielded such results; and check at test time that the results are attributable to the model's capabilities rather than issues with the test design. Findings from partner organizations and external evaluations of our models (or similar models) should also be incorporated into the final assessment, when available.

3. **Elicitation:** Demonstrate that, when given enough resources to extrapolate to realistic attackers, researchers cannot elicit sufficiently useful results from the model on the relevant tasks. We should assume that jailbreaks and model weight theft are possibilities, and therefore perform testing on models without safety mechanisms (such as harmlessness training) that could obscure these capabilities. We will also consider the possible performance increase from using resources that a realistic attacker would have access to, such as scaffolding, finetuning, and expert prompting. At minimum, we will perform basic finetuning for instruction following, tool use, minimizing refusal rates.

4. **Forecasting:** Make informal forecasts about the likelihood that further training and elicitation will improve test results between the time of testing and the next expected round of comprehensive testing.[8]

This testing and the subsequent capability decision should ideally be concluded within about a month of reaching the "notably more capable" threshold.

3.3. Capability Decision

If, after the comprehensive testing, we determine that the model is sufficiently below the relevant Capability Thresholds, then we will continue to apply the ASL-2 Standard.[9] The process for making such a determination is as follows:

- First, we will **compile a Capability Report** that documents the findings from the comprehensive assessment, makes an affirmative case for why the Capability Threshold is sufficiently far away, and advances recommendations on deployment decisions.

- The report will be **escalated to the CEO and the Responsible Scaling Officer**, who will (1) make the ultimate determination as to whether we have sufficiently established that we are unlikely to reach the Capability Threshold and (2) decide any deployment-related issues.

- In general, as noted in Sections 7.1.4 and 7.2.2, we will **solicit both internal and external expert feedback** on the report as well as the CEO and RSO's conclusions to inform future refinements to our methodology. For high-stakes issues, however, the CEO and RSO will likely solicit internal and external feedback on the report prior to making any decisions.

- If the CEO and RSO decide to proceed with deployment, they will **share their decision**—as well as the underlying Capability Report, internal feedback, and any external feedback—with the Board of Directors and the Long-Term Benefit Trust before moving forward.

[8] *Currently, these will be informal estimates of (1) the extent to which widely available elicitation techniques may improve and (2) how the model will perform on the same tasks when the next round of testing begins. As these are open research questions, we will aim to improve these forecasts over time so that they can be relied upon for risk judgments.*

[9] *In the case where the capability assessment shows a model is just barely below the threshold, the Responsible Scaling Officer may choose to limit further training to some amount less than the default 4x Effective Compute increase until ASL-3 measures are in place, in order to limit risk.*

If, however, we determine we are unable to make the required showing, we will act as though the model has surpassed the Capability Threshold.[10] This means that we will (1) upgrade to the ASL-3 Required Safeguards (see Section 4) and (2) conduct follow-up a capability assessment to confirm that the ASL-4 Standard is not necessary (see Section 5).

4. Safeguards Assessment

To determine whether the measures we have adopted satisfy the ASL-3 Required Safeguards, we will conduct a safeguards assessment. As noted, the Required Safeguards for each Capability Threshold are specified in Section 2. We will document our implementation of the Required Safeguards in a Safeguards Report.

4.1. ASL-3 Deployment Standard

When a model must meet the ASL-3 Deployment Standard, we will evaluate whether the measures we have implemented make us robust to persistent attempts to misuse the capability in question. To make the required showing, we will need to satisfy the following criteria:

1. **Threat modeling:** Make a compelling case that the set of threats and the vectors through which an adversary could catastrophically misuse the deployed system have been sufficiently mapped out, and will commit to revising as necessary over time.

2. **Defense in depth:** Use a "defense in depth" approach by building a series of defensive layers, each designed to catch misuse attempts that might pass through previous barriers. As an example, this might entail achieving a high overall recall rate using harm refusal techniques. This is an area of active research, and new technologies may be added when ready.

3. **Red-teaming:** Conduct red-teaming that demonstrates that threat actors with realistic access levels and resources are highly unlikely to be able to consistently elicit information from any generally accessible systems that greatly increases their ability to cause catastrophic harm relative to other available tools.[11]

4. **Rapid remediation:** Show that any compromises of the deployed system, such as jailbreaks or other attack pathways, will be identified and remediated promptly enough to prevent the overall system from meaningfully increasing an adversary's ability to cause catastrophic harm. Example techniques could include rapid vulnerability patching, the ability to escalate to law enforcement when appropriate, and any necessary retention of logs for these activities.

5. **Monitoring:** Prespecify empirical evidence that would show the system is operating within the accepted risk range and define a process for reviewing the system's performance on a reasonable cadence. Process examples include monitoring responses to jailbreak bounties, doing historical analysis or background monitoring, and any necessary retention of logs for these activities.

---

[10] *There may be a substantial period during which models are not demonstrably close to the Capability Threshold, but we nevertheless are unable to rule out the risk to our satisfaction, and thus choose to implement the Required Safeguards.*
[11] *This criterion does not attempt to specify the exact red-teaming protocol (e.g., number of hours, level of access, or pass-fail criteria). Setting a principled pass-fail threshold will depend on other factors, such as the quality of our monitoring and ability to respond to jailbreaks rapidly. Due to the likely ease of bypassing or removing safeguards via fine-tuning, it may be difficult or impossible for these red-teaming tests to pass if weights are released or if unmoderated fine-tuning access is provided to untrusted users.*

6. **Trusted users:** Establish criteria for determining when it may be appropriate to share a version of the model with reduced safeguards with trusted users. In addition, demonstrate that an alternative set of controls will provide equivalent levels of assurance. This could include a sufficient combination of user vetting, secure access controls, monitoring, log retention, and incident response protocols.

7. **Third-party environments:** Document how all relevant models will meet the criteria above, even if they are deployed in a third-party partner's environment that may have a different set of safeguards.

4.2. ASL-3 Security Standard

When a model must meet the ASL-3 Security Standard, we will evaluate whether the measures we have implemented make us highly protected against most attackers' attempts at stealing model weights.

We consider the following groups in scope: hacktivists, criminal hacker groups, organized cybercrime groups, terrorist organizations, corporate espionage teams, internal employees,[12] and state-sponsored programs that use broad-based and non-targeted techniques (i.e., not novel attack chains).

The following groups are out of scope for the ASL-3 Security Standard because further testing (as discussed below) should confirm that the model would not meaningfully increase their ability to do harm: state-sponsored programs that specifically target us (e.g., through novel attack chains or insider compromise) and a small number (~10) of non-state actors with state-level resourcing or backing that are capable of developing novel attack chains that utilize 0-day attacks.

To make the required showing, we will need to satisfy the following criteria:

1. **Threat modeling:** Follow risk governance best practices, such as use of the MITRE ATT&CK Framework to establish the relationship between the identified threats, sensitive assets, attack vectors and, in doing so, sufficiently capture the resulting risks that must be addressed to protect model weights from theft attempts. As part of this requirement, we should specify our plans for revising the resulting threat model over time.

2. **Security frameworks:** Align to and, as needed, extend industry-standard security frameworks for addressing identified risks, such as disclosure of sensitive information, tampering with accounts and assets, and unauthorized elevation of privileges with the appropriate controls. This includes:

    a. **Perimeters and access controls:** Building strong perimeters and access controls around sensitive assets to ensure AI models and critical systems are protected from unauthorized access. We expect this will include a combination of physical security, encryption, cloud security, infrastructure policy, access management, and weight access minimization and monitoring.

    b. **Lifecycle security:** Securing links in the chain of systems and software used to develop models, to prevent compromised components from being introduced and to ensure only trusted code and hardware is used. We expect this will include a combination of software inventory, supply chain security, artifact integrity, binary authorization, hardware procurement, and secure research development lifecycle.

[12] *We will implement robust insider risk controls to mitigate most insider risk, but consider mitigating risks from highly sophisticated state-compromised insiders to be out of scope for ASL-3. We are committed to further enhancing these protections as a part of our ASL-4 preparations.*

c. **Monitoring:** Proactively identifying and mitigating threats through ongoing and effective monitoring, testing for vulnerabilities, and laying traps for potential attackers. We expect this will include a combination of endpoint patching, product security testing, log management, asset monitoring, and intruder deception techniques.

d. **Resourcing:** Investing sufficient resources in security. We expect meeting this standard of security to require roughly 5-10% of employees being dedicated to security and security-adjacent work.

e. **Existing guidance:** Aligning where appropriate with existing guidance on securing model weights, including Securing AI Model Weights, Preventing Theft and Misuse of Frontier Models (2024); security recommendations like Deploying AI Systems Securely (CISA/NSA/FBI/ASD/CCCS/GCSB /GCHQ), ISO 42001, CSA's AI Safety Initiative, and CoSAI; and standards frameworks like SSDF, SOC 2, NIST 800-53.

3. **Audits:** Develop plans to (1) audit and assess the design and implementation of the security program and (2) share these findings (and updates on any remediation efforts) with management on an appropriate cadence. We expect this to include independent validation of threat modeling and risk assessment results; a sampling-based audit of the operating effectiveness of the defined controls; periodic, broadly scoped, and independent testing with expert red-teamers who are industry-renowned and have been recognized in competitive challenges.

4. **Third-party environments:** Document how all relevant models will meet the criteria above, even if they are deployed in a third-party partner's environment that may have a different set of safeguards.

4.3. Safeguards Decision

If, after the evaluations above, we determine that we have met the ASL-3 Required Safeguards, then we may proceed with deploying and training models above the Capability Threshold, provided we have also conducted a follow-up capability assessment. The process for determining whether we have met the ASL-3 Required Safeguards is as follows:

- First, we will **compile a Safeguards Report** for each Required Safeguard that documents our implementation of the measures above, makes an affirmative case for why we have satisfied them, and advances recommendations on deployment decisions.

- The Safeguards Report(s) will be **escalated to the CEO and the Responsible Scaling Officer**, who will (1) make the ultimate determination as to whether we have satisfied the Required Safeguards and (2) decide any deployment-related issues.

- In general, as noted in Sections 7.1.4 and 7.2.2, we will **solicit both internal and external expert feedback** on the report as well as the CEO and RSO's conclusions to inform future refinements to our methodology. For high-stakes issues, however, the CEO and RSO will likely solicit internal and external feedback on the report prior to making any decisions.

- If the CEO and RSO decide to proceed with deployment and training, they will **share their decision**—as well as the underlying Capability Report, internal feedback, and any external feedback—with the Board of Directors and the Long-Term Benefit Trust before moving forward.

- After the ASL-3 Required Safeguards are approved, they will be **revisited and re-approved at least annually** to re-affirm their suitability and sound implementation.

If, however, we are unable to make the showing required above, we will restrict model deployment and further scaling.

5. Follow-Up Capability Assessment

In parallel with upgrading a model to the ASL-3 Required Safeguards, we will conduct a follow-up capability assessment to determine that the model's capability falls sufficiently far away from the Capability Thresholds that would trigger ASL-4 Required Safeguards.

We will update this policy with the Capability Thresholds for the ASL-4 Required Safeguards. We are currently working on defining any further Capability Thresholds that would mandate ASL-4 Required Safeguards. Our update may not include all possible risks; we will prioritize capabilities that are likely to emerge earlier in frontier models.

As noted, before deploying any model that passes the Capability Thresholds for the ASL-3 Required Safeguards, we will conduct a capability assessment against the forthcoming Capability Thresholds for the ASL-4 Required Safeguards. We will follow the procedures outlined in Section 3.

6. Deployment and Scaling Outcomes

6.1. Continue Deployment and Further Scaling

To summarize the commitments and procedures outlined above, we may deploy or store a model if either of the following criteria are met: (1) the model's capabilities are sufficiently far away from the existing Capability Thresholds, making the current ASL-2 Standard appropriate; or (2) the model's capabilities have surpassed the existing Capabilities Threshold, but we have implemented the ASL-3 Required Safeguards and confirmed that the model is sufficiently far away from the next set of Capability Thresholds as to make the model ASL-3 Standard appropriate. We may also continue to train more capable models, conducting preliminary and comprehensive assessments as before.

6.2. Restrict Deployment and Further Scaling

In any scenario where we determine that a model requires ASL-3 Required Safeguards but we are unable to implement them immediately, we will act promptly to reduce interim risk to acceptable levels until the ASL-3 Required Safeguards are in place:

- **Interim measures:** The CEO and Responsible Scaling Officer may approve the use of interim measures that provide the same level of assurance as the relevant ASL-3 Standard but are faster or simpler to implement. In the deployment context, such measures might include blocking model responses, downgrading to a less-capable model in a particular domain, or increasing the sensitivity of automated monitoring.[13] In the security context, an example of such a measure would be storing the model weights in a single-purpose, isolated network that meets the ASL-3

[13] *When choosing amongst options that satisfy the safety criteria, we will implement whichever interim safeguards minimize changes to customer experience.*

Standard. In either case, the CEO and Responsible Scaling Officer will share their plan with the Board of Directors and the Long-Term Benefit Trust.

- **Stronger restrictions:** In the unlikely event that we cannot implement interim measures to adequately mitigate risk, we will impose stronger restrictions. In the deployment context, we will de-deploy the model and replace it with a model that falls below the Capability Threshold. Once the ASL-3 Deployment Standard can be met, the model may be re-deployed. In the security context, we will delete model weights. Given the availability of interim deployment and security protections, however, stronger restrictions should rarely be necessary.

- **Monitoring pretraining:** We will not train models with comparable or greater capabilities to the one that requires the ASL-3 Security Standard.[14] This is achieved by monitoring the capabilities of the model in pretraining and comparing them against the given model. If the pretraining model's capabilities are comparable or greater, we will pause training until we have implemented the ASL-3 Security Standard and established it is sufficient for the model. We will set expectations with internal stakeholders about the potential for such pauses.

7. Governance and Transparency

7.1. Internal Governance

To facilitate the effective implementation of this policy across the company, we commit to the following:

1. **Responsible Scaling Officer:** We will maintain the position of Responsible Scaling Officer, a designated member of staff who is responsible for reducing catastrophic risk, primarily by ensuring this policy is designed and implemented effectively. The Responsible Scaling Officer's duties will include (but are not limited to): (1) as needed, proposing updates to this policy to the Board of Directors; (2) approving relevant model training or deployment decisions based on capability and safeguard assessments; (3) reviewing major contracts (i.e., deployment partnerships) for consistency with this policy; (4) overseeing implementation of this policy, including the allocation of sufficient resources; (5) receiving and addressing reports of potential instances of noncompliance[15]; (6) promptly notifying the Board of Directors of any cases of noncompliance that pose material risk[16]; and (7) making judgment calls on policy interpretation[17] and application.

2. **Readiness:** We will develop internal safety procedures for incident scenarios. Such scenarios include (1) pausing training in response to reaching Capability Thresholds; (2) responding to a security incident involving model weights; and (3) responding to severe jailbreaks or vulnerabilities in deployed models, including restricting access in safety emergencies that cannot otherwise be mitigated. We will run exercises to ensure our readiness for incident scenarios.

---

[14] *We consider implementation of the ASL-3 Security Standard alone sufficient to continue training, regardless of whether the ASL-3 Deployment Standard is satisfied. "Comparable or greater capabilities" is operationalized as 1x or more in Effective Compute.*
[15] *In addition to noncompliance processes, we will (1) establish pathways for Anthropic staff to raise any issues related to this policy, including the overall risk levels of our models and implementation challenges; and (2) regularly review our compliance with this policy's procedural requirements.*
[16] *Cases deemed to present minimal additional risk may be reported to the Board in quarterly summary reports.*
[17] *In cases where this policy is unintentionally ambiguous, we will act in accordance with the Responsible Scaling Officer or CEO's judgment, and aim to clarify the ambiguity in the next policy update.*

3. **Transparency:** We will share summaries of Capability Reports and Safeguards Reports with Anthropic's regular-clearance staff, redacting any highly-sensitive information. We will share a minimally redacted version of these reports with a subset of staff, to help us surface relevant technical safety considerations.

4. **Internal review:** For each Capabilities or Safeguards Report, we will solicit feedback from internal teams with visibility into the relevant activities, with the aims of informing future refinements to our methodology and, in some circumstances, identifying weaknesses and informing the CEO and RSO's decisions.

5. **Noncompliance:** We will maintain a process through which Anthropic staff may anonymously notify the Responsible Scaling Officer of any potential instances of noncompliance with this policy. We will also establish a policy governing noncompliance reporting, which will (1) protect reporters from retaliation and (2) set forth a mechanism for escalating reports to one or more members of the Board of Directors in cases where the report relates to conduct of the Responsible Scaling Officer. Further, we will track and investigate any reported or otherwise identified potential instances of noncompliance with this policy. Where reports are substantiated, we will take appropriate and proportional corrective action and document the same. The Responsible Scaling Officer will regularly update the Board of Directors on substantial cases of noncompliance and overall trends.

6. **Employee agreements:** We will not impose contractual non-disparagement obligations on employees, candidates, or former employees in a way that could impede or discourage them from publicly raising safety concerns about Anthropic. If we offer agreements with a non-disparagement clause, that clause will not preclude raising safety concerns, nor will it preclude disclosure of the existence of that clause.

7. **Policy changes:** Changes to this policy will be proposed by the CEO and the Responsible Scaling Officer and approved by the Board of Directors, in consultation with the Long-Term Benefit Trust.[18] The current version of the RSP is accessible at www.anthropic.com/rsp. We will update the public version of the RSP before any changes take effect and record any differences from the prior draft in a change log.

7.2. Transparency and External Input

To advance the public dialogue on the regulation of frontier AI model risks and to enable examination of our actions, we commit to the following:

1. **Public disclosures:** We will publicly release key information related to the evaluation and deployment of our models (not including sensitive details). These include summaries of related Capability and Safeguards reports when we deploy a model[19] as well as plans for current and future

[18] *It is possible at some point in the future that another actor in the frontier AI ecosystem will pass, or be on track to imminently pass, a Capability Threshold without implementing measures equivalent to the Required Safeguards such that their actions pose a serious risk for the world. In such a scenario, because the incremental increase in risk attributable to us would be small, we might decide to lower the Required Safeguards. If we take this measure, however, we will also acknowledge the overall level of risk posed by AI systems (including ours), and will invest significantly in making a case to the U.S. government for taking regulatory action to mitigate such risk to acceptable levels.*

[19] *We currently expect that if we do not deploy the model publicly and instead proceed with training or limited deployments, we will likely instead share evaluation details with a relevant U.S. Government entity.*

comprehensive capability assessments and deployment and security safeguards.[20] We will also periodically release information on internal reports of potential instances of non-compliance and other implementation challenges we encounter.

2. **Expert input:** We will solicit input from external experts in relevant domains in the process of developing and conducting capability and safeguards assessments. We may also solicit external expert input prior to making final decisions on the capability and safeguards assessments.

3. **U.S. Government notice:** We will notify a relevant U.S. Government entity if a model requires stronger protections than the ASL-2 Standard.

4. **Procedural compliance review:** On approximately an annual basis, we will commission a third-party review that assesses whether we adhered to this policy's main procedural commitments (we expect to iterate on the exact list since this has not been done before for RSPs). This review will focus on procedural compliance, not substantive outcomes. We will also do such reviews internally on a more regular cadence.

[20] *These will be posted to www.anthropic.com/rsp-updates. We anticipate providing updates at least every 6-12 months. Where possible, we will include descriptions of the empirical evaluation results we believe would indicate that a model is no longer safe to store under the ASL-2 Standard. Our purpose in these updates is to provide sufficient detail to facilitate conversations about best practices for safeguards, capability evaluations, and elicitation.*

Appendices

Appendix A: Glossary

AI Safety Levels (ASLs)	Technical and operational standards for safely training and deploying frontier AI models. Higher ASLs correspond to stronger safety and security measures required for more capable models.
ASL-2 Standard	The current default standard for all Anthropic models, including security measures, safety testing, and automated misuse detection.
ASL-3 Standard	A higher level of safeguards required when a model cannot be certified as ASL-2 appropriate. It includes more stringent security and deployment measures designed to mitigate risks from more capable models.
Capability Report	A document attesting that a model is sufficiently far from each of the relevant Capability Thresholds, and therefore (still) appropriate for storing under an ASL-N Standard. It includes evaluation procedures, results, and other relevant evidence gathered around the time of testing.
Capability Thresholds	Specific AI capabilities that, if reached, would require stronger safeguards than the current baseline ASL-N standard provides.
Effective Compute	A scaling trend-based metric that accounts for both FLOPs and algorithmic improvements.
Evaluations	Empirical tests designed to provide early warning when a model is approaching a Capability Threshold. These tests are intended to trigger before a model actually reaches a dangerous capability.
FLOP(s)	Floating-Point Operation(s). The amount of computation required to train or run a model. The number of FLOPs can be used as one indicator of a model's computational complexity and, indirectly, its potential capabilities.
Long-Term Benefit Trust (LTBT)	Anthropic's Board of Directors approves the RSP and receives Capability Reports and Safeguards Reports. The LTBT is an external body that is consulted on policy changes and also provided with Capability Reports and Safeguards Reports. More details about the LTBT are available here.
Required Safeguards	The standard of safety and security measures that must be implemented when a model reaches a Capability Threshold.
Responsible Scaling Officer (RSO)	A designated staff member responsible for reducing catastrophic risk, primarily by ensuring this policy is designed and implemented effectively. Their duties include reviewing policy updates, approving reports, overseeing implementation, and approving deployments.
Safeguards Report	A document attesting that the implemented safeguards meet an ASL-N Standard. It details the design and planned implementation of safeguards, and evidence to demonstrate their expected effectiveness.

Appendix B: ASL-2 Standard

ASL-2 Deployment Standard:

1. **Acceptable use policies and model cards**: Publication of model cards for significant new models describing capabilities, limitations, evaluations, and intended use cases. Enforcement of a Usage Policy that restricts, at a minimum, catastrophic and high harm use cases, including using the model to generate content that could cause severe risks to the continued existence of humankind, or direct and severe harm to individuals.
2. **Harmlessness training and automated detection**: Training models to refuse requests to aid in causing harm, such as with Constitutional AI or other improved techniques, and the use of model enhanced trust and safety detection and enforcement.
3. **Fine-tuning protections:** In finetuning products, data is filtered for harmfulness, and models are subject to automated evaluation to check harmlessness features are not degraded. There are a very limited number of use cases where this tooling is disabled. These are negotiated on a case by case basis and considered only for extremely low risk use cases that involve company personnel.
4. **Vulnerability reporting channels**: Clearly indicated paths within the product for users to report harmful or dangerous model outputs, as well as a bug bounty for universal jailbreaks.

ASL-2 Security Standard: A security system that can likely thwart most opportunistic attackers.

1. **Supply chain:** Vendor and supplier security must be regularly reviewed to ensure that they meet security standards. Software updates should be frequently managed and compliance monitoring automated where possible.
2. **Offices:** Physical security should entail visitor access logs and restrictions protect on-site assets. Highly sensitive interactions should utilize advanced authentication like security keys. Network visibility should be maintained and office access controls and communications should maximize on-site protections.
3. **Workforce:** People-critical processes must represent a key aspect of cybersecurity. Mandatory periodic infosec training educates all employees on secure practices, like proper system configurations and strong passwords, and fosters a proactive "security mindset." Fundamental infrastructure and policies promoting secure-by-design and secure-by-default principles should be incorporated into the engineering process. An insider risk program should tie access to job roles. Rapid incident response protocols must be deployed.
4. **Compartmentalization:** Segmented system isolation must ensure limited blast radius. Features like zero trust architecture should require access from approved devices. Strict protocols must be deployed to regulate weight copies on company networks and limit storage to only approved, restricted systems.
5. **Infrastructure:** Standard security infrastructure, monitoring software, access management tools, and disk encryption provide a technology baseline. Process elements like incident reporting procedures, lost/stolen device protocols and Detection and Response should support these. External validation like SOC 2 compliance and continuous vulnerability management must ensure adaptations match infosec developments. Programs like bug bounties and vulnerability discovery should incentivize exposing flaws.
6. **Operations:** Ongoing configuration management, compliance drills, integrated security approaches and mandatory external reviews should embed security within regular operations and harden processes during organizational changes.

Appendix C: Detailed Capability Thresholds

This appendix contains detailed definitions for each Capability Threshold in Section 2.

Chemical, Biological, Radiological, and Nuclear (CBRN) weapons: The ability to significantly assist individuals or groups with basic STEM backgrounds in obtaining, producing, or deploying CBRN weapons.[21] We assess this by comparing what potential attackers could achieve with full model access versus 2023-level online resources, assuming they have funding and up to one year of time to invest, but no initial specialized expertise.[22]

Autonomous AI Research and Development: The ability to either: (1) Fully automate the work of an entry-level remote-only Researcher at Anthropic, as assessed by performance on representative tasks *or* (2) cause dramatic acceleration in the rate of effective scaling. Specifically, this would be the case if we observed or projected an increase in the effective training compute of the world's most capable model that, over the course of a year, was equivalent to two years of the average rate of progress during the period of early 2018 to early 2024. We roughly estimate that the 2018-2024 average scaleup was around 35x per year, so this would imply an actual or projected one-year scaleup of 35^2 = ~1000x.[23]

Model Autonomy checkpoint: The ability to perform a wide range of advanced software engineering tasks autonomously that could be precursors to full autonomous replication or automated AI R&D, and that would take a domain expert human 2-8 hours to complete. We primarily view this level of model autonomy as a checkpoint on the way to managing the risks of robust, fully autonomous systems with capabilities that might include (a) automating and greatly accelerating research and development in AI development (b) generating their own revenue and using it to run copies of themselves in large-scale, hard-to-shut-down operations.

[21] *We are uncertain how to choose a specific threshold, but we maintain a current list of specific CBRN capabilities of concern for which we would implement stronger mitigations. We treat these lists as sensitive, but we plan to share them with organizations such as AI Safety Institutes and the Frontier Model Forum, and keep these lists updated.*

[22] *This comparison is hard to make in practice; this note is to clarify the meaning of the conceptual threshold and the fact that this policy aims to measure risk relative to the world in 2023, so that we can understand how much risk the current generations of frontier models are creating.*

[23] *The 35x/year scaleup estimate is based on assuming the rate of increase in compute being used to train frontier models from ~2018 to May 2024 is 4.2 x/year (reference), the impact of increased (LLM) algorithmic efficiency is roughly equivalent to a further 2.8 x/year (reference), and the impact of post training enhancements is a further 3 x/year (informal estimate). Combined, these have an effective rate of scaling of 35 x/year.*

Changelog

September 19, 2023

RSP-2023 (aka RSP v1.0): Initial version, link here.

October 15, 2024

RSP-2024: This update introduces a more flexible and nuanced approach to assessing and managing AI risks while maintaining our commitment not to train or deploy models unless we have implemented adequate safeguards. Key improvements include new capability thresholds to indicate when we should upgrade our safeguards, refined processes for evaluating model capabilities and the adequacy of our safeguards (inspired by safety case methodologies), and new measures for internal governance and external input. We describe the most notable changes below.

ASL definition changed: The term "ASL" now refers to groups of technical and operational safeguards (it previously also referred to models). We also introduced the new concepts of Capability Thresholds and Required Safeguards. This change allows for more targeted application of safeguards based on specific capabilities, rather than broad model categories.

ARA threshold now a checkpoint: We replaced our previous autonomous replication and adaption (ARA) threshold with a "checkpoint" for autonomous AI capabilities. Rather than triggering higher safety standards automatically, reaching this checkpoint will prompt additional evaluation of the model's capabilities and accelerate our preparation of stronger safeguards. We previously considered these capabilities as a trigger for increased safeguards, motivated by an attempt to establish some threshold while we developed a better sense of potential threats. We now believe that these capabilities - at the levels we initially considered - would not necessitate the ASL-3 standard.

AI R&D threshold added: We added a new threshold for AI systems that can significantly advance AI development. Such capabilities could lead to rapid, unpredictable advances in AI, potentially outpacing our ability to evaluate and address emerging risks, and may also serve as an early warning sign for the ability to automate R&D in other domains.

Testing for Capability Thresholds: Rather than using prespecified evaluations, we now require an affirmative case that models are sufficiently far from Capability Thresholds. Predefined tests may miss emerging risks or be overly conservative relative to the actual threshold of concern. Our most accurate tests change frequently enough that it is more practical to use this new approach than to have our Board of Directors pre-approve evaluations.

Adjusted evaluation cadence: We adjusted the comprehensive assessment cadence to 4x Effective Compute or six months of accumulated post-training enhancements (this was previously three months). We found that a three-month cadence forced teams to prioritize conducting frequent evaluations over more comprehensive testing and improving methodologies.

Less prescriptive evaluation methodology: We have replaced some specifics in our previous testing methodology (e.g., using 1% of compute for elicitation or creating a 6x buffer), with more general requirements to (a) match expected efforts of potential adversaries and (b) provide informal estimates of how further scaling and research developments will impact model capabilities and performance on the same tasks. We have found that specific methodologies may become outdated

when new research developments are introduced. Although still an aspirational goal, the science of evaluations is not currently mature enough to make confident predictions about the precise buffer we should require between current models and a Capability Threshold.

More outcome-focused safeguard requirements: We have updated our ASL-3 safeguards requirements to be less prescriptive and more outcome-focused. Rather than detailing specific operational and technical safeguards, we now specify the overall security or deployment standards and requirements for meeting them. This is to allow us to adapt our safeguards more flexibly as our understanding of risks and possible safeguards improves.

Clarified ASL-3 and ASL-2 security threat models: We have clarified which actors are in and out of scope for the ASL-3 Security Standard. We also removed the commitment to protect against scaled attacks and distillation attacks from the ASL-2 Security standard. While distillation remains a concern for more capable models, models stored under ASL-2 safeguards have not yet reached potentially harmful Capability Thresholds.

Clarified requirements for deployments with trusted users: We have updated the ASL-3 Deployment Standard to allow for different levels of safeguards based on deployment context. For any general access systems, we still require passing intensive red-teaming. For internal use, safety testing and deployments to sufficiently trusted users, we will instead require a combination of access controls and monitoring.

New Capability and Safeguards Reports: We have introduced Capability Reports and Safeguard Reports. We expect that aggregating all the available evidence about model capabilities will provide decision makers with a more complete picture of the overall level of risk and improve our ability to solicit feedback on our work.

Internal and external accountability: We have made a number of changes to our previous "procedural commitments." These include expanding the duties of the Responsible Scaling Officer; adding internal critique and external expert input on capability and safeguard assessments; new procedures related to internal governance; and maintaining a public page for overviews of past Capability and Safeguard Reports, RSP-related updates, and future plans.