ANTHROP\C

Responsible Scaling Policy

Version 3.1

Effective April 2, 2026

For more information, see www.anthropic.com/responsible-scaling-policy

Contents

Introduction

Our Responsible Scaling Policy (RSP) is our voluntary framework for managing catastrophic risks from advanced AI systems. It establishes how we identify and evaluate risks, how we make decisions about AI development and deployment, and, from the perspective of the world at large, how we aim to make sure that the benefits of our models exceed their costs. We have always intended for our RSP to be a living document. We will continually update the RSP as we learn more about AI capabilities and risks, develop and refine technical safety measures, and gain more experience navigating an ecosystem in which the risks to society depend on the actions of many developers.

The major components of this third iteration are as follows:

Our **recommendations for industry-wide safety** outline what it would take, at an industry-wide level, to keep catastrophic risks reliably low through a period of rapid advances in AI capabilities. We lay this out in a table that maps capability thresholds to the mitigations we believe they call for. We also include our planned mitigations as a company, which are drawn from other sections of this policy and associated artifacts.

This approach represents a change from our previous RSP, driven by a collective action problem. The overall level of catastrophic risk from AI depends on the actions of multiple AI developers, not just one. Our previous RSP committed to implementing mitigations that would reduce our models' absolute risk levels to acceptable levels, without regard to whether other frontier AI developers would do the same. But from a societal perspective, what matters is the risk to the ecosystem as a whole. If one AI developer paused development to implement safety measures while others moved forward with training and deploying AI systems without strong mitigations, that could result in a world that is less safe—the developers with the weakest protections would set the pace, and responsible developers would lose their ability to do safety research and advance the public benefit. Although this situation has not yet arisen, it looks likely enough that we want to prepare for it.

We now separate our plans as a company—those which we expect to achieve regardless of what any other company does—from our more ambitious industry-wide recommendations. We aspire to advance the latter through a mixture of example-setting, addressing unsolved technical problems, advocacy through industry groups, and policy advocacy. But we cannot commit to following them unilaterally.

Frontier Safety Roadmaps are a new requirement under our RSP. These will describe our concrete plans for making progress across Security, Alignment, Safeguards, and Policy. Goals described in the Roadmaps are intended to be ambitious, yet achievable—providing the kind of forcing function that we consider to be a past success of our RSP. These are not hard commitments but rather public goals against which we will openly grade our progress.

Risk Reports are another new requirement. Risk Reports will provide detailed information on the safety profile of our models at the time of publication. They will go beyond describing model capabilities, addressing our thinking on how capabilities, threat models (the specific ways that models might pose threats), and active risk mitigations fit together, providing an assessment of the overall level of risk. These reports will reflect our reasoning as to whether we believe the risks of training or deploying our models are justified by their corresponding benefits to the world. They will be published online, with some redactions to protect sensitive details about, for example, our training methods and organizations with whom we work. As detailed below, we also aim to subject Risk Reports to review by credible, independent external parties.

Finally, our **governance** commitments are intended to promote internal and external accountability, similar to those in our previous RSP.

Our RSP is only one part of our overall approach to safety. For instance, although this policy focuses on catastrophic risks, they are not the only risks we consider important—our Usage Policy and societal impacts research address other concerns. Further, the RSP may serve some regulatory requirements, but it is not designed to be comprehensive. We want to keep it focused on our most central measures for addressing the

catastrophic risks we prioritize most, rather than expand it to address every obligation we face.[1] Where regulatory requirements exceed or differ from what the RSP covers, we will address them through separate documents.

1. Our Recommendations for Industry-Wide Safety

This section outlines our recommendations for what it would take, at an industry-wide level, to keep catastrophic risks reliably low through a period of rapid advances in AI capabilities. We lay this out in a three-column table. The left column identifies capability thresholds that would call for heightened mitigations. The middle column provides an overview of our planned mitigations, which we have set forth more fully in our Frontier Safety Roadmap and other sections of this policy. The right column describes our recommendations for industry-wide safety at each threshold.

The distinction between our plans as a company (middle column) and our industry-wide recommendations (right column) reflects the limitations of any single AI developer's ability to ensure safety across the industry. In particular, we cannot unilaterally and unconditionally commit to staying in line with the industry-wide recommendations in the right column. However, these recommendations will drive important aspects of our work:

- We use these recommendations as the north star for our risk mitigations planning as well as our public policy work. We will strive to advance these recommendations through a mixture of example-setting, addressing unsolved technical problems, advocacy through industry groups, and policy advocacy.
- We have also adopted a set of competitor-contingent commitments (see Appendix A) aimed at staying in line with these recommendations in scenarios where we can be confident that other relevant AI developers are doing the same.

At this point in AI's rapid development, we cannot presently give highly specific advance detail on what evaluations will determine whether risk thresholds have been passed, or what risk mitigations will be needed to achieve safety. Our recommendations for industry-wide safety are thus structured around requiring analysis and arguments making a strong case for safety, rather than AI Safety Levels (more in Appendix B). This leaves flexibility in how risk thresholds are evaluated and how safety is achieved and argued for. But it creates a challenge: one actor's view of what constitutes good risk assessment and mitigation may be very different from another's.

[1] "Catastrophic risk" as used in our RSP refers generally to risks of the most severe potential harms from advanced AI, such as existential threats or fundamental destabilization of global systems. We use this term in its plain meaning rather than adopting any specific statutory definition. Where laws such as California SB 53 define this or similar terms with specific thresholds, we address those requirements in separate compliance frameworks.

Ultimately, the best way for these recommendations to be implemented is likely via governance of all relevant frontier AI developers by third parties that determine which developers need to provide risk analyses and make arguments for the safety of their systems, and determine which such arguments are adequate.[2] To the extent this takes the form of national regulation, different countries should attempt to harmonize their governance, including standards of evidence, to avoid a race to the bottom. In the shorter run, independent bodies (standards-setting organizations, auditors, etc.) might review such arguments and enforce high quality standards for private AI developers via voluntary mechanisms.

We expect that the recommendations for industry-wide safety will evolve significantly, as we learn more about AI capabilities, threat models, and risk mitigations. We hope these recommendations will become increasingly specific over time.

[2] Note that an AI developer might provide a single analysis covering a broad category of models, such that there isn't a new analysis needed for each release. Additionally, determinations of which AI developers need to provide such analysis should be made with an effort to spare smaller AI developers from unnecessary compliance burden.

Capability or usage threshold	Mitigations—our plan as a company[3]	Mitigations—ambitious industry-wide recommendations
Non-novel chemical/biological weapons production. AI systems with the ability to significantly help individuals or groups with basic technical backgrounds (e.g., undergraduate STEM degrees) create/obtain and deploy chemical and/or biological weapons with serious potential for catastrophic damages.	We will maintain or improve on our ASL-3 protections, which include classifier guards at least as robust as our initial Constitutional Classifiers; access controls for trusted users with exemptions to classifier guards; red-teaming, bug bounties, and threat intelligence for continually assessing the threat of jailbreaks; and a number of noteworthy security controls. Specifics may change, but we will maintain equally or more robust measures over time and will publish updates in our Risk Reports. We expect to continuously meet the criteria in the right column, although we cannot make guarantees about an evolving landscape with continually adaptive attackers.	A frontier developer should make a strong argument that individual users and relatively small teams will not become significantly more likely to cause catastrophic harm via their usage of product surfaces or via theft of model weights. This will likely require: • Restrictions on model behavior, and/or measures for quickly detecting and acting on Usage Policy violations, accompanied by a strong case that these measures are difficult to reliably, sustainably circumvent via jailbreaking. • Precautions against opportunistic theft of model weights, such as centralized controls on third-party applications and software updates.

[3] This column summarizes commitments drawn from other sections of this policy and associated artifacts. It should be read as an overview, not as indicating any separate or additional plans.

Capability or usage threshold	Mitigations—our plan as a company[3]	Mitigations—ambitious industry-wide recommendations
Novel chemical/biological weapons production. AI systems with the ability to significantly help threat actors (for example, moderately resourced expert-backed teams) create/obtain and deploy chemical and/or biological weapons with potential for catastrophic damages far beyond those of past catastrophes such as COVID-19.	We will apply protections at least as strong as our ASL-3 protections (see previous row) to an expanded set of potential use cases for AI, covering the most likely vectors for this threat. Additionally, we will identify the most concerning specific threat pathways, create policy recommendations for early detection and response for such threats, and share this content with policymakers.	A frontier developer should make a strong argument that threat actors will not become significantly more likely to cause the sort of catastrophic harm discussed in the lefthand column via their usage of product surfaces or via theft of model weights. This will likely require similar measures to those from the previous row, but to a higher standard—to the point where even well-resourced and -staffed threat actors would be unlikely to reliably jailbreak models or cause catastrophic harm via unauthorized access to or modification of models (including via stolen or modified model weights). This would likely mean security roughly in line with RAND SL4, but it depends on the capabilities of the strongest and most plausible threat actors that are not bound by a credible governance regime enforcing the recommendations for industry-wide safety outlined here.
High-stakes sabotage opportunities. AI systems that are highly relied on and have extensive access to sensitive assets as well as moderate capacity for autonomous, goal-directed operation and subterfuge—such that it is plausible these AI systems could (if directed toward this goal, either deliberately or inadvertently) carry out sabotage leading to irreversibly and substantially higher odds of a later global catastrophe. In the near term, this possibility will likely be most applicable to AI systems that are extensively used within major AI companies, with the opportunity to manipulate how their successor systems are trained and deployed as well as the evidence used to assess their	We will detail the state of our AI systems' capabilities and propensities, our monitoring practices, and the overall level of risk in our Risk Reports. We expect to continually be able to meet the criteria in the right column, although we cannot make guarantees about an evolving technology that may increasingly have the ability to detect and manipulate testing.	A frontier developer should make a strong argument that AI systems will not carry out sabotage leading to irreversibly and substantially higher odds of a later global catastrophe. This case may initially be relatively simple and rely heavily on capability limitations, if it is first required when the risk is merely plausible. As risk becomes harder to rule out, this case will likely include some combination of: • Internal compartmentalization, restriction, and code review to prevent excessive sabotage opportunities for AI models. • Capability assessments demonstrating that AI models lack the ability to carry out irreversible (which would generally mean unnoticed) sabotage. • Monitoring and/or restricting AI behavior and usage internally. • Evidence that AI models lack the propensity to deceive, manipulate, or sabotage users.

Capability or usage threshold	Mitigations—our plan as a company[3]	Mitigations—ambitious industry-wide recommendations
safety. Down the line, this possibility may come to apply to AI systems deployed within government and other high-stakes settings.		
Automated R&D in key domains. AI systems that can fully automate, or otherwise dramatically accelerate, the work of large, top-tier teams of human researchers in domains where fast progress could cause threats to international security and/or rapid disruptions to the global balance of power—for example, energy, robotics, weapons development and AI itself. For now, our evaluations will focus specifically on AI R&D, as this domain likely plays to AI systems' current strengths and is more tractable to assess than capabilities in other domains. Additionally, AI R&D alone could cause acceleration in AI capabilities improvements, to the point where all of the threats listed above (and more) develop very quickly. We will consider this threshold to be met if we determine that either (1) our models would be able to fully substitute for our entire set of Research Scientists and Research Engineers, at competitive costs	We will: • Resource and complete significant "moonshot R&D for security" projects, to explore ambitious and possibly unconventional ways to achieve unprecedented levels of security against the world's best-resourced attackers. • Achieve an "eyes on everything" state for our internal AI development. We will comprehensively gather, centralize, and maintain logs for all critical AI-development activities, and use AI to analyze them for issues including security threats, concerning behavior by insiders (humans as well as AI systems themselves), and training processes or data that are out of line with the public Constitution that shapes and defines our AI models. • Perform systematic	A frontier developer should make a strong argument that: • No user or team of users (including those backed by top-tier states) will become significantly more likely to cause catastrophic harm via their usage of product surfaces or via theft of model weights. This will likely require similar measures to those listed in row 1, but to a higher standard, to the point where even well-resourced and -staffed threat actors would be unlikely to reliably jailbreak models or cause catastrophic harm via unauthorized access to or modification of models (including via stolen or modified model weights). Accomplishing this would likely mean security roughly in line with RAND SL4. Security requirements would be calibrated to defend against the strongest plausible threat actors who are not bound by a credible industry-wide safety regime. Actors subject to such a regime would not need to be treated as threats to each other's model weights. • Even malicious employees and other insiders with maximal levels of access will not be significantly enabled to cause catastrophic harm. This requires (among other things) accounting for internal tools that are less restricted than product surfaces, and for the possibility of unauthorized modification of models. This will likely require an internal Usage Policy and strong internal compartmentalization, controls and/or monitoring to restrict the ability of employees and contractors (up to and including the company's CEO as well as its most privileged technical employees) to circumvent the Usage Policy. • AI models have not been deliberately or inadvertently trained with dangerous goals, or are otherwise unlikely to autonomously cause

Capability or usage threshold	Mitigations—our plan as a company[3]	Mitigations—ambitious industry-wide recommendations
(i.e., within a factor of 5); or (2) there is "dramatic acceleration" of the pace of AI progress for reasons that likely relate to the automation of AI R&D. We would consider scenario (2) to have occurred where (a) we observe or expect *double the rate of progress*[4] *in AI aggregate capabilities* compared to the rate we'd expect in the absence of significant AI contributions to AI R&D and (b) it is plausible that this doubling is substantially attributable to the automation of research and/or engineering (as opposed to other factors, such as increased headcount, compute, or general productivity), such that continuation of the trend in AI progress could lead to even greater acceleration. This capability threshold is intended to reflect our definition of *highly capable* models (see Section 3.6). It may be sensible to add earlier, and/or easier-to-measure, thresholds that trigger less demanding versions of the mitigations for this threshold.	alignment assessments to examine Claude's behavioral patterns and propensities, meaningfully incorporating mechanistic interpretability and adversarial red-teaming to test our auditing methods. • Develop our internal red-teaming of our deployment safeguards to the point where our internal red-teaming performs better at finding potential jailbreaks than the collective abilities of the participants in our established bug bounty programs. • Publish Risk Reports with the status of, and noteworthy findings from, all of the above, subject to external review by at least one expert, experienced, credible, candid and disinterested external party. We broadly expect to lead the industry in practices that reduce the risks from AI, although we cannot unilaterally make guarantees about the safety level	catastrophic harm. This will likely require similar measures to those listed above under "High-stakes sabotage opportunities" (some combination of internal compartmentalization, restriction and code review; monitoring AI behavior; and evidence that AI models lack the propensity to deceive and manipulate users), but to a greater degree. In particular, at this point AI systems might be responsible for much of the research and analysis that underpins risk assessment, and might have strong capabilities for deception, manipulation and obfuscation of evidence, in which case analyses of threats from AIs should follow very high evidentiary standards with thorough and careful analysis of the possibility that much of the key evidence is suspect due to the possibility of manipulation by AI systems.

[4] "Double the rate of progress" means "as much progress in one year as one would see in two years at baseline." For example, if baseline progress involved a 3x scaleup in compute and a 3x improvement in algorithmic efficiency (for a 9x "effective scaleup"), "double the rate of progress" would entail something like an 81x effective scaleup. This is not the same idea as "doubling researchers' productivity," since doubling inputs does not necessarily double the rate of progress.

Capability or usage threshold	Mitigations—our plan as a company[3]	Mitigations—ambitious industry-wide recommendations
	of AI this advanced.	

2. Frontier Safety Roadmap

We will maintain a Frontier Safety Roadmap laying out ambitious but achievable goals for improving our risk mitigations—charting our progress as a company and raising the bar over time. Maintaining and reporting on this Roadmap is part of our work under the RSP. It will be shared with all full-time employees as well as our Board of Directors (Board) and Long-Term Benefit Trust (LTBT), and published in redacted form. We will provide updates on whether we achieve the goals, and set new goals when we do.

Our Frontier Safety Roadmap is subject to change. Some changes may simply reflect our evolving understanding of how best to mitigate key risks. However, we will strive to avoid situations where we revise the goals in a less ambitious direction simply because we are unable to achieve them. By establishing this expectation, we hope to create a forcing function for work that would otherwise be challenging to appropriately prioritize and resource, as it requires collaboration (and in some cases sacrifices) from multiple parts of the company and can be at cross-purposes with immediate competitive and commercial priorities. Publishing our Frontier Safety Roadmap may also help inform broader industry and policy discussions on AI safety.

Our current Frontier Safety Roadmap is available at anthropic.com/responsible-scaling-policy/roadmap. We will also keep past Roadmaps available at that link.

3. Risk Reports

We will publish **Risk Reports** discussing the risks of our systems and how we have made determinations about whether to continue AI development and deployment in light of the risks. These have significant content in common with system cards, but we are adding additional structure and process aimed at presenting our overall assessments of risk.

3.1. Scope and Timing

Scope. A Risk Report will cover all publicly deployed models at the time of its publication. It will also cover internally deployed models when we determine that these models could pose significant risks[5] beyond those posed by models that are covered by a prior Risk Report. While there are a variety of reasons we might classify an internal model this way, this will—at a minimum—include any internal models that we are deploying for large-scale, fully autonomous research.

Models fitting the above description are abbreviated below as "*in-scope models*." We may also voluntarily include additional models in a Risk Report, e.g., to contribute to general discourse, but such inclusion does not expand the commitments below.

Timing. We will publish a Risk Report every 3-6 months. Note that unlike system cards, Risk Reports will not be published with each new model release. Additionally:

- When we publicly deploy a model that we determine is significantly more capable than any of the models covered in the most recent Risk Report, we will publish a discussion (in our System Card or elsewhere) of how that model's capabilities and propensities affect or change analysis in the Risk Report.
- Within 30 days of determining that we have an internally deployed model that is in-scope (per the description above), we will publish a discussion (in a System Card or elsewhere) of how that model's capabilities and propensities affect or change the analysis in the Risk Report.

[5] Specifically, risks arising from the capability thresholds in our recommendations for industry-wide safety (see Section 1).

3.2. General Expectations for Risk Reports

Several principles guide how we approach Risk Reports:

- We intend for our Risk Reports to be direct, candid, and informative about how we see the risks of our systems and our state of preparedness for them.
- In particular, we will acknowledge when we view certain models as posing significant risks in absolute terms, even if our marginal contribution to overall ecosystem risk may be relatively limited when taking other developers' AI models into account.
- We will put significant effort into investigating (for example, via capability evaluations) the case for risk, and into sharing what we find.
- When a Risk Report describes risk mitigations we have in place (or plan to implement shortly), we will keep our future practices in line with this description *or* track and report noteworthy changes and deviations (generally via subsequent Risk Reports). We should make a strong attempt to ensure that ongoing (as opposed to temporary) changes do not significantly increase the level of risk.

3.3. Contents

Factual information. We will describe how we identify, evaluate, and mitigate catastrophic risks. A Risk Report will document the following:

1. **Threat model identification:** Our criteria for determining which catastrophic risks (i.e., threat models) we assess.
2. **Threat model specification:** The relevant threat models (which will, at a minimum, include those discussed above).
3. **Evidence (including evaluations) about relevant model capabilities and behaviors:** For each in-scope model, capability and alignment evaluations (conducted internally and by external parties as appropriate), their results, and (as appropriate) other evidence we considered in assessing the level of risk.
4. **Risk mitigations:** For each in-scope model, the mitigations we are implementing across security, deployment safeguards, and alignment domains, along with discussion of their effectiveness.
5. **Additional relevant factors:** Any other considerations material to our risk analysis.

Risk analyses. We will provide our reasoning and conclusions regarding both specific threat models and overall risk levels. Our analyses will include:

1. **Threat-specific risk assessment:** For each threat model, we will analyze remaining absolute risk–i.e., the leftover risk after accounting for our mitigations. We will also discuss whether we believe we've crossed relevant thresholds in our recommendations for industry-wide safety (see Section 1), and (as relevant) whether we believe we're meeting the risk mitigation standards corresponding to them.
2. **Overall risk assessment:** We will provide an overall risk assessment.
3. **Risk-benefit determination:** We will explain whether, and if so why, we believe the identified risks are justified by corresponding benefits.
4. **Looking forward:** We will outline our plans for continuing to monitor and mitigate the relevant risks over time.

Review of past Risk Reports and decisions. We will address:

1. **Changes in risk mitigation practices**—noteworthy cases in which our risk mitigation practices deviated (including temporarily) or changed from what we discussed in our previous Risk Report over the relevant period, and their implications for the overall level of risk.

2. **Decisions to internally deploy in-scope models** that would not otherwise be reviewed in a Risk Report (because they happened in between Risk Reports). We will discuss how these decisions were made and whether they appear reasonable in light of any noteworthy new information that has come to light since then.

3. **Changes to our Frontier Safety Roadmap and any cases where we failed to meet our goals.**

Marginal risk and ecosystem analysis. If we determine that the absolute level of risk being imposed industry-wide by AI systems such as ours is high, and are justifying our decision to move forward based partly on a marginal risk analysis,[6] we will additionally address the following:

1. **Competitive landscape analysis:** What we know about how our current and future model capabilities and risk mitigations compare to those of relevant competitors.

2. **Role in risk assessment:** How these comparisons factored into our risk assessment.

3. **Benefits analysis:** The benefits, if any, of continuing to maintain our existing models or of advancing frontier capabilities, including considerations related to our mission and our ability to contribute to AI safety research and policy development.

4. **Advocacy efforts:** The steps we took to raise public awareness of the relevant risks and to encourage appropriate regulatory action, including our engagement with policymakers and other AI developers.

We will conduct the assessments above with respect to each in-scope model. To avoid redundancy, we will likely analyze similar models collectively rather than individually, with appropriate justification for any such groupings.

3.4. Procedures

1. **Initial assessment and drafting:** Our internal subject matter experts will conduct risk assessments and draft the Risk Report.

2. **Review and feedback:** We will solicit comprehensive internal feedback on the report, focusing on identifying potential methodological weaknesses, analytical gaps, or areas requiring additional evidence or clarification. We may also use this feedback to improve or refine the report itself. We will usually also seek feedback from trusted external parties with relevant expertise.

3. **Executive approval:** The Risk Report, along with the internal feedback and any available external feedback, will be sent to the CEO and Responsible Scaling Officer (RSO) for final review and approval. The CEO and RSO will make the ultimate determination regarding the adequacy of the risk assessment and any downstream deployment or development plans.

4. **Governance notification:** Following approval of a Risk Report, the CEO and RSO will promptly share their decision(s), the underlying Risk Report, and internal feedback with both the Board and the LTBT.

5. **Modified process when marginal risk analysis is important to our case.** In the event that marginal risk analysis (see previous section) plays a major role in a decision to move forward, explicit approval of the Risk Report by the Board and LTBT (rather than just the CEO and RSO) will be required.

---

[6] By "*marginal risk analysis*," we mean arguing that the risks imposed by our systems in particular are relatively lower when keeping in mind the risks unavoidably posed by other AI systems.

3.5. Publication and Redactions

We will publish a public version of our Risk Report. We will aim to minimize redactions to the public version of the report. Reasons we may redact material include but are not limited to:

1. **Legal compliance:** To comply with legal obligations on disclosing information, such as export control regulations, national security restrictions, or contractual obligations with third parties.
2. **Intellectual property protection:** To protect proprietary information, including technical details of our models and training methodologies, trade secrets, or other confidential business information.
3. **Public safety considerations:** To protect public safety by not disclosing information that could be exploited to cause harm.
4. **Privacy:** To protect the personal privacy of individuals.

3.6. External Review

We will work toward a practice of seeking comprehensive, public external review on our Risk Reports. This means working with one or more third-party organizations that will receive private versions of our Risk Reports (unredacted or with minimal redactions, as discussed below) and publish comprehensive commentary on them. Commentary will address topics including the quality of our reasoning, the validity of our risk assessments, the overall level of risk, and whether the redactions we've made for the public version are reasonable and appropriate.

Our intent is for external reviewers' judgments to carry significant weight in the eyes of the public as well as our employees. But there are no well-established organizations or procedures for this sort of practice, and we are approaching it as an experiment.

At a minimum, we will complete a full external review process (described below) with at least one external reviewer anytime a Risk Report covers *highly capable* models and is *significantly redacted,* defined as follows:

- A model is "*highly capable*" if we conclude that it crosses the threshold for automated AI R&D described in Section 1.
- "*Significantly redacted*" means that the redactions omit information a reasonable external safety researcher would consider important in evaluating the overall level of risk, such that a reader of only the public version could not meaningfully assess whether they agree with our conclusions. A report shall be deemed significantly redacted if, in the judgment of the RSO, CEO, Board, or LTBT, it meets this description. Note that unredacted reports will be shared with a large number of employees, who will be in a position to raise concerns if they believe the public version meets this description.

3.6.1. Selecting external reviewers

In consultation with the Board and LTBT, we will select external reviewers that:

- Have significant experience and expertise regarding evaluations for dangerous AI capabilities and propensities. It's particularly important that they be knowledgeable about potential ways such evaluations might be misleading (for example, alignment faking).
- Have reputational and other incentives making them likely to be candid about their assessment of risks, rather than focused on writing comments that Anthropic will approve of. For example, external review parties should not be teams whose revenue, reputation and success depend entirely on Anthropic and similar companies continuing to work with them.

- Do not have conflicts of interest with respect to Anthropic. At a minimum, a reviewing organization itself may not have a financial interest in Anthropic; and the individuals involved in conducting the review, as well as anyone above them in the reporting chain within their organization, may not have a financial interest in Anthropic or close personal relationships with anyone at Anthropic (i.e., family relationships, romantic relationships, or shared living arrangements).

3.6.2. Timing and access

The external review process will involve sharing the Risk Report with one or more external reviewers within one week of submitting the same report to our Board and LTBT. We will ask the external reviewers to provide public commentary on our report within 30 days of receipt. We will try to work toward a process that involves the full external review being completed prior to Board/LTBT review (and may require this later).

For purposes of external review, the only redactions to the Risk Report will be those necessary to comply with legal prohibitions or to maintain our legal rights.[7] We expect that we will also invest some time in answering follow-up questions from parties doing external review.

3.6.3. Contents of external review

The external review will address:

1. **Adequacy of information:** Whether the Risk Report contains sufficient information to assess the identified risks;
2. **Analytical rigor:** The strength of the Risk Report's reasoning and analysis;
3. **Areas of disagreement:** Whether the external reviewer disagrees with any of the Risk Report's key claims and, if so, the reasons for any such disagreements; and
4. **Risk reduction recommendations:** Recommendations for further reducing identified risks.

The external reviewer will also evaluate our redaction decisions while avoiding disclosure of the redacted content itself. In particular, the review will cover:

1. **Redaction scope:** The general nature and scope of redactions;
2. **Redaction justification:** Whether the external reviewer generally agrees or disagrees with the publicly stated reasons for the redactions and, if relevant, the reasons for any disagreements;
3. **Balancing test:** Whether the redactions strike a reasonable balance between Anthropic's legitimate interests and society's interest in transparency; and
4. **Materiality:** Whether any of the redactions in the public report are material to any of the external reviewer's substantive disagreements with the report's claims.

Public comments. We will ask the external reviewer to memorialize its findings on the topics above in a written report that is made public. External reviewers will be bound by obligations not to disclose confidential information (e.g., confidential intellectual property, matters of national security, or proprietary details such as model architecture, cost, or size), but beyond that will not be restricted in what they can publish, including concerns about the Risk Report or Anthropic's conduct in connection with the external review.

[7] For example, we may redact information if sharing it would endanger our *legal right to treat it as confidential.* This would be an uncommon situation. It does not mean we would redact information merely because it is sensitive or to reduce the risk of a leak.

4. Governance

We commit to the following governance measures to promote internal and external accountability.

1. **Responsible Scaling Officer:** We will maintain the position of RSO, a designated member of staff who is responsible for the implementation of this policy. The RSO's duties will include (but are not limited to): (1) as needed, proposing updates to this policy; (2) approving relevant model development or deployment decisions based on our risk assessments; (3) reviewing major contracts (e.g., deployment partnerships) for consistency with this policy; (4) overseeing the implementation of this policy, including the allocation of sufficient resources; (5) receiving and addressing reports of potential instances of noncompliance; and (6) making judgment calls on policy interpretation and application.

2. **Internal transparency:** We will share final, unredacted Risk Reports with Anthropic's regular-clearance staff.

3. **Noncompliance reporting:** We will maintain a process for Anthropic staff to submit anonymous or identified reports regarding potential noncompliance with this policy. Staff will have more than one option for who receives these reports, including the RSO, and at least one executive who does not report to the RSO. When we receive a report, we will promptly investigate, take appropriate and proportional corrective action if it is substantiated, and document the report and our findings. We will provide quarterly updates to the Board regarding reports of potential noncompliance, whether substantiated or not. If we determine that a report is (1) substantiated and (2) involves a material safety risk, we will promptly notify the Board and we may provide public notice of the same. Finally, we will protect reporters from retaliation, and where a report concerns the conduct of the RSO, at least one recipient will be a member of the Board.

4. **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.

5. **Internal review:** We will regularly conduct an internal review of our implementation of this policy.

6. **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. This review will focus on procedural compliance, not substantive outcomes.

7. **Policy changes:** Changes to the RSP will be proposed by the CEO and RSO, and approved by the Board in consultation with the LTBT. If we update the RSP, we will publicly share the updated version prior to or on its effective date and will record the differences from the prior draft in the Change Log. We will maintain the current version of the RSP on our website.

Appendices

Appendix A: Commitments Related to Competitors

These commitments are necessarily high-level and limited. In many cases, we will not have enough information to determine that the relevant scenario applies and will have to use our best judgment to deal with uncertainty. But to the extent that other relevant AI developers prioritize safety and invest in legible demonstrations that they are doing so—as we intend to—commitments like this may help avoid an inadvertent "race to the bottom" on safety. Further, the commitments below do not preclude us from taking cautionary action, such as refraining from training or deploying models, in other circumstances. Mitigating the risks from our models is a top priority for us, and we would strongly consider pausing development and/or deployment to improve the safety profiles of our models even in cases not covered below.

Scenario	Commitment
Anthropic in the lead. We have developed or will imminently develop a *highly capable*[8] model; and we have clear evidence that no other competitor will soon develop such a model.	We will require a strong argument that catastrophic risk is contained, along the lines of our recommendations for industry-wide safety (see Section 1). We will delay AI development and deployment as needed to achieve this, until and unless we no longer believe we have a significant lead.
Competitors have strong safety measures. We have strong evidence that all competitors who have developed, or will soon develop, a *highly capable* frontier model are able to make strong arguments that catastrophic risk is contained, in the spirit of our recommendations for industry-wide safety (see Section 1).	For our *highly capable* frontier models, we will meet or exceed the overall risk reduction posture of these competitors, as far as we can tell based on our best efforts to assess that posture. Until we are able to do so, we will delay AI development and deployment as needed to achieve this.
General upleveling. We have strong evidence that a competitor has implemented a risk mitigation that: (1) represents a significant improvement on reduction of our prioritized risks relative to our analogous mitigations; and (2) we could implement at comparable (or lower) effort or cost to our competitor.	We will make a significant effort to meet or exceed that performance standard. However, we will not necessarily delay AI development and deployment in this scenario.

Appendix B: Notes on ASLs

Earlier editions of our RSP defined "AI Safety Levels" with specific lists of required controls. We still use this concept to refer to, and distinguish between, *present* levels of risk mitigations—those that we maintain for existing AI models. (For example, our initial Risk Report uses this distinction.) However, when defining the risk mitigations needed for *future* levels of AI capability, we have found that providing a specific list of controls is overly rigid, and we instead prefer to focus on what sort of argument an AI developer should make (and what sorts of actors it should address) regarding the risk level from its systems.

Changelog

September 19, 2023 (RSP v1.0)

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

October 15, 2024 (RSP v2.0)

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

[8] This and other italicized instances of "highly capable" use the term as defined in Section 3.6.

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 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.

March 31, 2025 (RSP v2.1)

RSP-2025: This update clarifies which Capability Thresholds would require enhanced safeguards beyond our current ASL-3 standards. The key changes include:

New Capability Thresholds: 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. We have also 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 provided additional detail on the corresponding Required Safeguards.

Iterative Commitment: We have adopted a general commitment to reevaluate our Capability Thresholds whenever we upgrade to a new set of Required Safeguards. We have decided not to maintain a commitment to define ASL-N+1 evaluations by the time we develop ASL-N models; such an approach would add unnecessary complexity because Capability Thresholds do not naturally come grouped in discrete levels. We believe it is more practical and sensible instead to commit to reconsidering the whole list of Capability Thresholds whenever we upgrade our safeguards.

May 14, 2025 (RSP v2.2)

ASL-3 Security: This update excludes both sophisticated insiders and state-compromised insiders from the ASL-3 Security Standard. Previously, only "highly sophisticated state-compromised insiders" were explicitly excluded. The model capabilities and threat models corresponding with the ASL-3 Security Standard do not warrant protection against either group: the CBRN-3 threat models entail large numbers of users having access to unguarded models (which is more likely to occur through a universal jailbreak than via model theft), and the relatively small number of employees who might be capable of model theft does not significantly affect the risk level. For AI R&D-4, the threat models generally do not depend on model weight theft and instead entail AI systems engaging in autonomous internal sabotage.

February 24, 2026 (RSP v3.0)

This update is a comprehensive rewrite of our RSP. For a summary of changes and the thinking behind them, see here.

April 2, 2026 (RSP v3.1)

This revision addresses the following points: (1) how we operationalize the Automated R&D capability threshold; (2) how we use internal feedback on Risk Reports; and (3) that we may consider pausing development or deployment even where the commitments described in Appendix A are not triggered. These changes are mostly clarificatory in nature; we don't see them as significant changes to substance. Change (1) reflects further discussion of our operationalization of the capability threshold, and involves some substantive changes to make it better reflect our underlying intent. This update also includes minor edits for style or clarity.