Anthropic's Frontier Safety Roadmap

We believe that AI capabilities will improve rapidly in the coming years. We will need to quickly and dramatically improve our state of preparedness in a number of areas, especially:

Security

Preventing theft, sabotage and/or manipulation of our AI models.

Safeguards

Preventing dangerous use of our models via product surfaces and within Anthropic itself.

Alignment

Ensuring that our models themselves do not autonomously cause harm, and instead consistently behave in line with our Constitution.

Policy

Laying out and advocating for a tangible path for policymakers to effectively manage AI risks across the industry, worldwide.

Our Frontier Safety Roadmap aims to chart a course in public for some of our highest-priority goals. Our hopes are twofold.

First, many (though not all) of these goals will need large company-wide efforts. Many people, across many different departments, will need to coordinate with each other across reporting lines. By clearly and publicly articulating these goals, we are

ANTHROP\C

Try Claude

Overview Goals Expectations

Anthropic's Frontier Safety Roadmap

We believe that AI capabilities will improve rapidly in the coming years. We will need to quickly and dramatically improve our state of preparedness in a number of areas, especially:

Security

Preventing theft, sabotage and/or manipulation of our AI models.

Safeguards

Preventing dangerous use of our models via product surfaces and within Anthropic itself.

Alignment

Ensuring that our models themselves do not autonomously cause harm, and instead consistently behave in line with our Constitution.

Policy

Laying out and advocating for a tangible path for policymakers to effectively manage AI risks across the industry, worldwide.

Our Frontier Safety Roadmap aims to chart a course in public for some of our highest-priority goals. Our hopes are twofold.

First, many (though not all) of these goals will need large company-wide efforts. Many people, across many different departments, will need to coordinate with each other across

aiming for a "forcing function" that helps us pull off this kind of coordination and prioritization. We believe our Responsible Scaling Policy has served this function effectively in the past, as with our ASL-3 protections.

Second, we want to share how we're thinking about the future of safety, at the level of tangible practices. We hope that other AI developers will take inspiration from these goals, and publish their own that we can learn from in turn. And we think policymakers and customers should be aware of where AI safety practices are headed.

These roadmaps are 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 because we simply can't execute.

We will provide updates on whether we achieve the goals, and set new goals whenever we do.

Note: we have made redactions from our public Frontier Safety Roadmap for reasons including protecting sensitive IP and not giving too much information about our current protections to threat actors. The unredacted version is shared with all full-time employees as well as our board and Long-Term Benefit Trust (LTBT).

Updates

You can find an archive of all our previous Frontier Safety Roadmap goals here.

1. **April 2, 2026: Launching "moonshot R&D" projects.** We completed this goal by selecting and beginning work on two projects, detailed below.
2. **April 2, 2026: Principles for data retention.** We completed our goal of writing a "comprehensive internal report to identify how our Safeguards could be improved by updating our data retention policies." The report was completed and sent to leadership on March 29, 2026. As stated below, within 6 weeks of that date, we will publish a new goal related to this or announce that we are not doing so.
3. **May 5, 2026: Principles for data retention.** We removed this content from the page after deciding not to set a new goal on this front. Based in part on the analysis we completed as a prior roadmap goal, we are undertaking a major project to improve our data retention practices for certain future models from a safety point of view, and we will be putting out public content explaining our thinking. We don't feel that setting a separate date-bound goal in this context would be helpful as a forcing function for the goal—we feel that other internal processes will provide adequate forcing functions.

reporting lines. By clearly and publicly articulating these goals, we are aiming for a "forcing function" that helps us pull off this kind of coordination and prioritization. We believe our Responsible Scaling Policy has served this function effectively in the past, as with our ASL-3 protections.

Second, we want to share how we're thinking about the future of safety, at the level of tangible practices. We hope that other AI developers will take inspiration from these goals, and publish their own that we can learn from in turn. And we think policymakers and customers should be aware of where AI safety practices are headed.

These roadmaps are 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 because we simply can't execute.

We will provide updates on whether we achieve the goals, and set new goals whenever we do.

Note: we have made redactions from our public Frontier Safety Roadmap for reasons including protecting sensitive IP and not giving too much information about our current protections to threat actors. The unredacted version is shared with all full-time employees as well as our board and Long-Term Benefit Trust (LTBT).

Updates

You can find an archive of all our previous Frontier Safety Roadmap goals here.

1. **April 2, 2026: Launching "moonshot R&D" projects.** We completed this goal by selecting and beginning work on two projects, detailed below.
2. **April 2, 2026: Principles for data retention.** We completed our goal of writing a "comprehensive internal report to identify how our Safeguards could be improved by updating our data retention policies." The report was completed and sent to leadership on March 29, 2026. As stated below, within 6 weeks of that date, we will publish a new goal related to this or announce that we are not doing so.
3. **May 5, 2026: Principles for data retention.** We removed this content from the page after deciding not to set a new goal on this front. Based in part on the analysis we completed as a prior roadmap goal, we are undertaking a major project to improve our data retention practices for certain future models from a safety point of view, and we

4. **May 5, 2026: "Moonshot R&D" projects.** We changed our target date for Phase 1 from May 15, 2026 to September 30, 2026 because we decided to focus the relevant resources on accelerating our "Leveling up across the board" goal for the time being. While we still don't expect to *fully complete* the latter goal in the next few months, we believe we will make significant progress on it, and that the safety benefits of doing so are greater than what we'd realize with our original prioritization. As for the latter goal, our initial discussions have left us skeptical that isolated networks are a feasible path forward inside the next 1-2 years, when powerful AI may be developed. We still plan to complete Phase 1, which will hopefully leave us with a more confident and grounded view.
5. **July 8, 2026: A roadmap for policymakers.** We consider our Advanced AI Framework to fulfill this goal. We plan to publish a new policy goal to replace it.
6. **July 10, 2026: Upholding Claude's Constitution.** We clarified the language on when the public Constitution must be updated. Where we previously stated we planned to update it within 90 days of relevant internal updates, we now specify that we will update it to match the most recent trained-on Constitution within 90 days of relevant internal or external deployments of models trained on updated Constitutions.
7. **July 29, 2026: "Moonshot R&D" projects.** Corrected a date typo in the main text. The Phase 1 deadline is September 30, 2026 (per the goal set May 5, 2026 seen above), but we had erroneously stated September 15.

Overview **Goals** Expectations

Our goals as of July 10th, 2026

Collaborations to rapidly and responsibly develop AI alongside measures to mitigate the potential for our systems to cause undue harm.

All | Security | Safeguards | Alignment | Policy | Cross-cutting

---

Security Target: September 30, 2026

"Moonshot R&D" projects

will be putting out public content explaining our thinking. We don't feel that setting a separate date-bound goal in this context would be helpful as a forcing function for the goal—we feel that other internal processes will provide adequate forcing functions.

4. **May 5, 2026: "Moonshot R&D" projects.** We changed our target date for Phase 1 from May 15, 2026 to September 30, 2026 because we decided to focus the relevant resources on accelerating our "Leveling up across the board" goal for the time being. While we still don't expect to *fully complete* the latter goal in the next few months, we believe we will make significant progress on it, and that the safety benefits of doing so are greater than what we'd realize with our original prioritization. As for the latter goal, our initial discussions have left us skeptical that isolated networks are a feasible path forward inside the next 1-2 years, when powerful AI may be developed. We still plan to complete Phase 1, which will hopefully leave us with a more confident and grounded view.
5. **July 8, 2026: A roadmap for policymakers.** We consider our Advanced AI Framework to fulfill this goal. We plan to publish a new policy goal to replace it.
6. **July 10, 2026: Upholding Claude's Constitution.** We clarified the language on when the public Constitution must be updated. Where we previously stated we planned to update it within 90 days of relevant internal updates, we now specify that we will update it to match the most recent trained-on Constitution within 90 days of relevant internal or external deployments of models trained on updated Constitutions.
7. **July 29, 2026: "Moonshot R&D" projects.** Corrected a date typo in the main text. The Phase 1 deadline is September 30, 2026 (per the goal set May 5, 2026 seen above), but we had erroneously stated September 15.
8. **October 19, 2026: "Moonshot R&D" projects.** We believe we broadly achieved the substantive components of the first goal. We created a prototype of what our key workflows and infrastructure would look like if they were subject to extreme security practices. We've explored the longest lead-time systems that needed to change to achieve air-gapped operations. Most of our effort has been spent building prototype environments in which we run workloads to test what productivity cost they'll incur and did not focus on commensurate physical security controls, in favor of other work. We plan to deploy inference systems in isolated environments in the near future. We also created an inventory of the needed controls and as the next step, have begun putting them into production in our environments. We believe we broadly achieved the second goal in full. For the provable inference project, we built a prototype that runs on production inference and checks that outputs come from a specific set of model weights. In developing this prototype, we rolled out signing on all of our internal token traffic, which has already caught one extremely rare, benign integrity glitch. We

Security is an ongoing and immediate priority for us, but it is also a long-term challenge where we'll need to be creative and explore promising and incomplete ideas. This is because we may, at some point in the future, be targets of the world's best-resourced attackers. Our moonshot R&D projects are exploring ambitious, possibly unconventional ways to achieve unprecedented levels of security.

Read more +

Security Target: July 1, 2027

Leveling up across the board

While innovation and creativity are important to security, it's also the case that security is fundamentally about the strength of the weakest link. That means we need to get a large number of small things right. We have a list of key security improvements that would represent a significant leveling up across many areas of

now use the term "inference verification," rather than "provable inference", our original aim with this moonshot is to achieve high assurance in the integrity of our token stream, which can be achieved in several ways, and "provable inference" suggests a narrower approach than we believe is necessary.

9. **October 19, 2026: Leveling up across the board.** We added goals in two areas that matter more as our own datacenter buildout and bare-metal deployments ramp up: datacenter physical security and platform security.
10. **October 19, 2026: World-class internal red-teaming.** We completed this goal ahead of its January 1, 2027 target. We built an automated red-teaming system and compared it with the bug bounties we ran against our biological weapons classifiers between April 20 and September 11, 2026. On the same 8 bounty questions and 3 generations of production classifiers, it matched or beat the best human submission in all 24 question-classifier pairs, winning 23. It did this in 24 to 48 hours, compared with weeks to months of bounty time. We judge these findings to be more substantial in aggregate than our bug bounty findings, which is the test we set. We plan to publish a more in-depth blog post on this in the future. We removed this goal from the roadmap, as we have completed it.
11. **October 19, 2026: Upholding Claude's Constitution.** We have kept our Constitution up to date and public. Our automated alignment assessment evaluates alignment with the Constitution as a core metric, and this is reported for all assessed models. Interpretability methods have become an important part of our overall alignment assessment methodology, and evidence produced using interpretability is featured in all recent system cards. We have red-teamed our alignment arguments using model organisms and published the findings in Risk Reports, though not as extensively as we had intended when drafting this goal. We initially ran an oversight process over a representative sample of our post-training data and rewards to check alignment with the Constitution, but found that it surfaced nothing beyond what our existing review processes were already catching, so we no longer run it routinely. A recent re-run on a sample of training data again flagged nothing. We are retiring this goal in its current form while we spend more time reflecting on which goals are most effective for alignment, and we will publish more about our thinking in the future. At a minimum, we plan to continue to update the public Constitution to match the most recent trained-on Constitution within 90 days of relevant internal or external deployments of models trained on updated Constitutions.
12. **October 19, 2026: New policy goals.** We added two new policy goals. These are a policy proposal on embedded third-party evaluators (due December 15, 2026) and a policy framework for targeting the pace of AI development (due January 15, 2027).

security, and we will aim to achieve a strong majority of them (details below). This will be a large, wide-reaching effort to achieve major hardening of our internal systems against attempts to compromise our model weights, model training, etc.

Our goals as of October 19th, 2026

Collaborations to rapidly and responsibly develop AI alongside measures to mitigate the potential for our systems to cause undue harm.

Security Target: July 1, 2027

Leveling up across the board

While innovation and creativity are important to security, it's also the case that security is fundamentally about the strength of the weakest link. That means we need to get a large number of small things right. We have a list of key security improvements that would represent a significant leveling up across many areas of security, and we will aim to achieve a strong majority of them (details below). This will be a large, wide-reaching effort to achieve major hardening of our internal systems against attempts to compromise our model weights, model training, etc.

Read more +

Safeguards Target: January 1, 2027

World-class internal red-teaming

Today, we continually test our safeguards against AI misuse by red-teaming them in a variety of ways, including via a bug bounty program that offers rewards to anyone who can identify jailbreak techniques of concern. In the future, we may need to prevent misuse from actors more sophisticated than what we can simulate with these methods. We believe we can develop a method for red-teaming our systems (likely involving significant automation) that outperforms the collective contributions from the hundreds of participants in our bug bounty. This could put us in position to continue leveling up over time and stay ahead of even state-sponsored threat actors with large teams devoted to bypassing our safeguards.

Read more +

Safeguards Target: January 1, 2027

Fully automated attack investigations

We have already seen coordinated, sophisticated efforts to misuse our systems for espionage. Stopping these efforts requires more than blocking individual AI outputs - it can involve dedicated investigation of patterns across many users and interactions. We believe we can build a system that conducts effective investigations with minimal or no human involvement, a major step forward for our ability to detect misuse at scale. Our initial efforts will focus on investigating potential cyber attacks on a subset of product surfaces.

Read more +

Alignment Target: October 1, 2026

Upholding Claude's Constitution

We recently published Claude's Constitution, which explains the kind of entity we want Claude to be and what principles we hope its behavior will follow. We already have a number of measures in place to ensure that the way we train Claude is in line with the Constitution, but we aim to make these measures more complete and systematic:

- We will keep our public Constitution up to date and run a systematic oversight process over our training data to evaluate alignment with it.
- We will perform systematic "alignment assessments" to examine Claude's behavioral patterns and propensities, and evaluate whether they are in line with the spirit of the Constitution. These alignment assessments will incorporate findings from our interpretability research, and will validate the effectiveness of our methods using testing exercises with intentionally misaligned models. We will publish our findings in our system cards or Risk Reports.

Read more +

Safeguards Target: January 1, 2027

Fully automated attack investigations

Read more +

Cross-cutting Target: January 1, 2027

Moving toward an "eyes on everything" state for our internal AI development activities

For many reasons - preventing theft or sabotage of our models, ensuring adherence to Claude's Constitution, and ensuring that our model training hasn't been manipulated by insiders or attackers (or AI models themselves) - we want our internal environment to be intensively monitored and comprehensively understood. We aim to establish comprehensive, centralized records of all our critical AI development activities, and to use AI to analyze these records for issues including concerning behavior by insiders (both human and AI) and security threats.

We have already seen coordinated, sophisticated efforts to misuse our systems for espionage. Stopping these efforts requires more than blocking individual AI outputs - it can involve dedicated investigation of patterns across many users and interactions. We believe we can build a system that conducts effective investigations with minimal or no human involvement, a major step forward for our ability to detect misuse at scale. Our initial efforts will focus on investigating potential cyber attacks on a subset of product surfaces.

Read more +

Cross-cutting Target: January 1, 2027

Moving toward an "eyes on everything" state for our internal AI development activities

For many reasons - preventing theft or sabotage of our models, ensuring adherence to Claude's Constitution, and ensuring that our model training hasn't been manipulated by insiders or attackers (or AI models themselves) - we want our internal environment to be intensively monitored and comprehensively understood. We aim to establish comprehensive, centralized records of all our critical AI development activities, and to use AI to analyze these records for issues including concerning behavior by insiders (both human and AI) and security threats.

Read more +

Policy

A roadmap for policymakers update

We recently completed our “roadmap for policymakers” goal and expect to publish a new goal in the coming weeks.

Expectations as AI capabilities advance

With the above goals in mind, this section outlines the risk mitigations we believe we’ll be in position to implement as AI capabilities advance, focused on the threat models we emphasize in our Responsible Scaling Policy.

Today’s capabilities and mitigations February 22nd, 2026

Our most powerful models today—those with the ability to significantly help individuals or groups with basic technical backgrounds create/obtain and deploy

Read more +

Policy Target: December 15, 2026

Policy proposal on embedded third-party evaluators

Independent evaluation is a core element of our Responsible Scaling Policy and our policy framework for mitigating catastrophic AI risks. We recently committed to going further by embedding independent evaluators, giving them deeper access to models and internal information than external evaluators typically have, while preserving their independence.

We will develop and publish a policy proposal on how governments can support deeper access for independent evaluators, including how evaluators are funded, what access they need, and how their independence is protected. We will draw on our own experience working with them.

Policy Target: January 15, 2027

Policy framework for targeting the pace of AI development

Our current policy framework focuses on blocking or deterring the deployment of dangerous AI systems. However, we believe policymakers may also need to target the *pace* of AI development directly, because the automation of AI R&D could accelerate progress faster than governance or safety measures can keep up. Speed itself could become an important risk factor.

We are working on additional policy proposals that provide options for policymakers to address the risks raised by the speed of AI development, particularly under conditions of recursive self-improvement. We will either publish a proposal or publish an explanation of why we decided not to.

Alignment

chemical and/or biological weapons with serious potential for catastrophic damages—are safeguarded with ASL-3 protections.

We plan to maintain or improve these protections for relevant models. The key elements of these protections are: safeguards at least as robust as our initial Constitutional Classifiers; access controls for trusted users with exemptions to classifier guards; and methods such as red-teaming, bug bounties, and threat intelligence for continually assessing the threat of jailbreaks; and a number of noteworthy security controls. The specific tools and methods we use may evolve, but our protections will be at least as rigorous as those we have in place today.

We hope and expect that these measures will remain sufficient to meet our goal of being both (a) robust to persistent attempts to misuse a potentially dangerous capability, and (b) highly protected against most attackers' attempts at stealing model weights. We will continue to adapt our defenses as attack methods evolve. However, while we can and will maintain the rigor of our protections, we cannot assure a specific level of effectiveness against future attackers, given attackers' continuous adaptability and the evolving risk landscape.

Additionally, today's models show strong and improving capabilities for extended, autonomous technical work, and are increasingly relied on for writing high-stakes code. In our Risk Reports, we now regularly analyze the possibility that these models might execute high-stakes sabotage. We believe that our current alignment assessment methods and monitoring practices are sufficient to make the case that the risk of catastrophic sabotage is very low (though not negligible).

Potential advances in the next few months

As capabilities improve, we are monitoring for significant advances in our models' chemical and biological weapons capabilities, focusing on the point where the models could significantly help threat actors (for example, moderately resourced expert-backed teams) create or obtain and deploy chemical and/or biological weapons with potential for catastrophic damages far beyond those of past catastrophes in this category such as COVID-19.

If and when we determine that this is the case, we will apply protections at least as strong as our current ASL-3 protections (see above) to an expanded set of potential use cases for AI, covering the most likely vectors for this threat. Additionally, we will

Update on "Upholding Claude's Constitution"

We are retiring this goal in its current form while we spend more time reflecting on which goals are most effective for alignment, and we will publish more about our thinking in the future. At a minimum, we plan to continue to update the public Constitution to match the most recent trained-on Constitution within 90 days of relevant internal or external deployments of models trained on updated Constitutions. See more on our updates page.

Expectations as AI capabilities advance

With the above goals in mind, this section outlines the risk mitigations we believe we'll be in position to implement as AI capabilities advance, focused on the threat models we emphasize in our Responsible Scaling Policy.

Today's capabilities and mitigations February 22nd, 2026

Our most powerful models today—those with the ability to significantly help individuals or groups with basic technical backgrounds create/obtain and deploy chemical and/or biological weapons with serious potential for catastrophic damages—are safeguarded with ASL-3 protections.

We plan to maintain or improve these protections for relevant models. The key elements of these protections are: safeguards at least as robust as our initial Constitutional Classifiers; access controls for trusted users with exemptions to classifier guards; and methods such as red-teaming, bug bounties, and threat intelligence for continually assessing the threat of jailbreaks; and a number of noteworthy security controls. The specific tools and methods we use may evolve, but our protections will be at least as rigorous as those we have in place today.

identify the most concerning specific threat pathways, create policy recommendations for early detection and response for such threats, and share this content with policy leaders.

Automated R&D

We believe it is plausible, as soon as early 2027, that our AI systems could 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.

By that point, we expect to have accomplished most of the goals listed above, including:

- Resourcing and completing significant "moonshot R&D security" projects, currently focused on provable inference and a mock secure research environment.
- Developing world-class internal red-teaming and automated attack investigations.
- Publishing and advocating for a roadmap for policymakers to achieve industry-wide safety without unnecessarily limiting the benefits from AI development or slowing the AI development of democracies relative to that of autocracies.
- Achieving an "eyes on everything" state for our internal AI development.
- Consistently implementing systematic alignment audits and other measures for upholding Claude's Constitution.

We hope and expect to meet these targets on time or ahead of schedule, set new ones, and continually raise the bar for our risk mitigations.

We hope and expect that these measures will remain sufficient to meet our goal of being both (a) robust to persistent attempts to misuse a potentially dangerous capability, and (b) highly protected against most attackers' attempts at stealing model weights. We will continue to adapt our defenses as attack methods evolve. However, while we can and will maintain the rigor of our protections, we cannot assure a specific level of effectiveness against future attackers, given attackers' continuous adaptability and the evolving risk landscape.

Additionally, today's models show strong and improving capabilities for extended, autonomous technical work, and are increasingly relied on for writing high-stakes code. In our Risk Reports, we now regularly analyze the possibility that these models might execute high-stakes sabotage. We believe that our current alignment assessment methods and monitoring practices are sufficient to make the case that the risk of catastrophic sabotage is very low (though not negligible).

Potential advances in the next few months

As capabilities improve, we are monitoring for significant advances in our models' chemical and biological weapons capabilities, focusing on the point where the models could significantly help threat actors (for example, moderately resourced expert-backed teams) create or obtain and deploy chemical and/or biological weapons with potential for catastrophic damages far beyond those of past catastrophes in this category such as COVID-19.

If and when we determine that this is the case, we will apply protections at least as strong as our current ASL-3 protections (see above) 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 policy leaders.

Automated R&D

We believe it is plausible, as soon as early 2027, that our AI systems could 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.

By that point, we expect to have accomplished most of the goals listed above, including:

- Resourcing and completing significant “moonshot R&D security” projects, currently focused on provable inference and a mock secure research environment.
- Developing world-class internal red-teaming and automated attack investigations.
- Publishing and advocating for a roadmap for policymakers to achieve industry-wide safety without unnecessarily limiting the benefits from AI development or slowing the AI development of democracies relative to that of autocracies.
- Achieving an “eyes on everything” state for our internal AI development.
- Consistently implementing systematic alignment audits and other measures for upholding Claude’s Constitution.

We hope and expect to meet these targets on time or ahead of schedule, set new ones, and continually raise the bar for our risk mitigations.

Government

Healthcare

Higher education

K-12 teachers

Legal

Life sciences

Nonprofits

Sales

Small business

Claude Platform

Overview

Developer docs

Pricing

Ecosystem

Marketplace

Regional compliance

Claude on AWS

Google Cloud

Microsoft Foundry

Console login

Resources

Blog

Claude partner network

Community

Connectors

Courses

Customer stories

Developer blog

Help and security

Terms and policies

Our mission

Anthropic is a Public Benefit Corporation. Our mission is to ensure the world safely makes the transition through transformative AI.

Log in

Claude.ai ↗

Claude console ↗

© 2026 Anthropic PBC

Engineering at Anthropic

Events

Plugins

Powered by Claude

Service partners

Tutorials

Use cases

Programs

Startups

Scientists

Help and security

Availability

Status

Support center

Company

Anthropic

Careers

Leadership

Policy

Economic Futures

Research

News

Claude's Constitution

Claude Corps

Keep thinking

Policy on the AI Exponential

Responsible Scaling Policy

Security and compliance

Transparency

Terms and policies

Privacy choices

Privacy policy

Consumer health data privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Terms of Service: US K-12

Data Processing Agreement: US K-12

Usage policy

© 2026 Anthropic PBC. Services in the EU are provided by Anthropic Ireland Limited.

Revisions Summary Report

Frontier_Safety_Roadmap___Anthropic__former_.pdf compared with
Frontier_Safety_Roadmap___Anthropic.pdf
2026-11-04 13:15:19
144 changes

Added 24 Removed 96 Replaced 24 Moved 0 Formatting 0

#1 (p.1)	Added (Text)	Try Claude
#2 (p.2)	Added (Text)	
#3 (p.2)	Added (Text)	
#4 (p.2)	Added (Text)	
#5 (p.2)	Added (Text)	
#6 (p.2)	Added (Text)	
#7 (p.3)	Added (Text)	
#8 (p.3)	Added (Text)	
#9 (p.3)	Added (Text)	
#10 (p.3)	Added (Text)	
#11 (p.3)	Added (Text)	
#12 (p.3)	Added (Text)	
#13 (p.3)	Added (Text)	
#14 (p.3)	Replaced (Text)	**Before:** Overview Goals Expectations **After:** 8. October 19, 2026: “Moonshot R&D”projects.We believe we broadly achieved the substantive components of the first goal. We created a prototype of what our key workflows and infrastructure would look like if they were subject to extreme security practices. We’ve explored the longest lead-time systems that needed to change to achieve air-gapped operations. Most of our effort has been spent building prototype environments in which we run workloads to test what productivity cost they’ll incur and did not focus on commensurate physical security controls, in favor of other work. We plan to deploy inference systems in isolated environments in the near future. We also created an inventory of the needed controls and as the next step, have begun putting them into production in our environments. We believe we broadly achieved the second goal in full. For the provable inference project, we built a prototype that runs on production inference and checks that outputs come from a specific set of model weights. In developing this prototype, we rolled out signing on all of our internal token traffic, which has already caught one extremely rare, benign integrity glitch. We

#15 (p.3 → p.4)	Replaced (Text)	**Before:** Our goals as of July 10th, 2026 Collaborations to rapidly and responsibly develop AI alongside measurestomitigatethepotential for our systems to cause undue harm. **After:** now use the term “inference verification,” rather than “provable inference”, our original aim with this moonshot is to achieve high assurance in the integrity of our token stream, which can be achieved in several ways, and “provable inference” suggests a narrower approach than we believe is necessary. 9. October 19, 2026:
#16 (p.3)	Removed (Text)	Security Target: September 30, 2026
#17 (p.3)	Removed (Text)	"Moonshot R&D"projects
#18 (p.3)	Removed (Text)	All Security Safeguards Alignment Policy Cross-cutting
#19 (p.4)	Removed (Text)	Security is an ongoing and immediate priority for us, but it is also a long-term challenge where we’ll need to be creative and explore promising and incomplete ideas. This is because we may, at some point in the future, be targets of the world’s best-resourced attackers. Our moonshot R&D projects are exploring ambitious, possiblyunconventionalwaysto achieve unprecedented levels of security.
#20 (p.4)	Removed (Text)	Read more
#21 (p.4)	Removed (Text)	Security Target: July 1, 2027

#22 (p.4)	Replaced (Text)	**Before:** While innovation and creativity are important to security, it's also the case that security is fundamentally about the strength of the weakest link. That means we needto get a large number of small things right. We have a list of key security improvements that would represent a significant leveling up across many areas of **After:** . We added goals in two areas that matter more as our own datacenter buildout and bare-metal deployments ramp up: datacenter physical security and platform security. 10. October 19, 2026: World-class internal red-teaming. We completed this goal ahead of its January 1, 2027 target. We built an automated red-teaming system and compared it with the bug bounties we ran against our biological weapons classifiers between April 20 and September 11, 2026. On the same 8 bounty questions and 3 generations of production classifiers, it matched or beat the best human submission in all 24 question-classifier pairs, winning 23. It did this in 24 to 48 hours, compared with weeks to months of bounty time. We judge these findings to be more substantial in aggregate than our bug bounty findings, which is the test we set. We plan to publish a more in-depth blog post on this in the future. We removed this goal from the roadmap, as we have completed it. 11. October 19, 2026: Upholding Claude's Constitution. We have kept our Constitution up to date and public. Our automated alignment assessment evaluates alignment with the Constitution as a core metric, and this is reported for all assessed models. Interpretability methods have become an important part of our overall alignment assessment methodology, and evidence produced using interpretability is featured in all recent system cards. We have red-teamed our alignment arguments using model organisms and published the findings in Risk Reports, though not as extensively as we had intended when drafting this goal. We initially ran an oversight process over a representative sample of our post-training data and rewards to check alignment with the Constitution, but found that it surfaced nothing beyond what our existing review processes were already catching, so we no longer run it routinely. A recent re-run on a sample of training data again flagged nothing. We are retiring this goal in its current form while we spend more time reflecting on which goals are most effective for alignment, and we will publish more about our thinking in the future. At a minimum, we plan to continue to update the public Constitution to match the most recent trained-on Constitution within 90 days of relevant internal or external deployments of models trained on updated Constitutions. 12. October 19, 2026: New policy goals. We added two new policy goals. These are a policy proposal on embedded third-party evaluators (due December 15, 2026) and a policy framework for targeting the pace ofAI development (due January 15, 2027).
#23 (p.5)	Replaced (Text)	**Before:** security, and we will aim to achieve a strong majority of them (details below). This will be a large, wide-reaching effort to achieve major hardening of our internal systems against attempts to compromise our model weights, model training, etc. **After:** Overview Goals Expectations
#24 (p.7 → p.5)	Replaced (Text)	**Before:** Read more **After:** Our goals as of October 19th, 2026 Collaborations to rapidly and responsibly develop AI alongside measurestomitigatethepotential for our systems to cause undue harm.

#25 (p.7 → p.5)	Replaced (Text)	**Before:** Safeguards Target: January 1, 2027 **After:** All Security Safeguards Alignment Policy Cross-cutting
#26 (p.8 → p.5)	Replaced (Text)	**Before:** World-class internal red-teaming Today, we continually test our safeguards against AI misuse by red-teaming them in a varietyofways, including via a bug bounty program that offers rewards to anyone who can identify jailbreak techniques of concern. In the future, we may need to preventmisuse from actors more sophisticated than what we can simulate with thesemethods. We believe we can develop a method for red-teaming our systems (likely involving significant automation) that outperforms the collective contributions from the hundreds of participants in our bug bounty. This could put us in position to continue leveling up over time and stay ahead of even state-sponsored threat actors with large teams devoted to bypassing our safeguards. **After:** Security Target: July 1, 2027
#27 (p.5)	Added (Text)	Leveling up across the board
#28 (p.5)	Added (Text)	While innovation and creativity are important to security, it's also the case that security is fundamentally about the strength of the weakest link. That means we need to get a large number of small things right. We have a list of key security improvements that would represent a significant leveling up across many areas of security, and we will aim to achieve a strong majority of them (details below). This will be a large, wide-reaching effort to achieve major hardening of our internal systems against attempts to compromise our model weights, model training, etc.
#29 (p.11)	Added (Text)	
#30 (p.11)	Added (Text)	
#31 (p.11)	Added (Text)	
#32 (p.9 → p.11)	Replaced (Text)	**Before:** Alignment Target: October 1, 2026 **After:** Cross-cutting Target: January 1, 2027
#33 (p.9 → p.11)	Replaced (Text)	**Before:** Upholding Claude's Constitution We recently published **After:** Moving toward an "eyes on everything"state for our internal AI development activities For many reasons - preventing theft or sabotage of our models, ensuring adherence to
#34 (p.9 → p.11)	Replaced (Text)	**Before:** which explains the kind of entity we want Claude to be and what principles we hope its behavior will follow. We already have a number of measures in place to ensure that the way we train Claude is in line withthe Constitution, but we aim to make these measures more complete and systematic: **After:** and ensuring that our model training hasn't been manipulated by insiders or attackers (or AI models themselves)- we want our internal environment to be intensively monitored and comprehensively understood. We aim to establish comprehensive, centralized records of all our critical AI development activities, and to use AI to analyze these records for issues including concerning behavior by insiders (both human and AI) and security threats.

#35 (p.9 → p.14)	Replaced (Text)	**Before:** We will keep our public Constitution up to date and run a systematic oversight processoverourtraining data to evaluate alignment with it. We will perform systematic “alignment assessments” to examine Claude’s behavioral patterns and propensities, and evaluate whether they are in line with thespiritofthe Constitution. These alignment assessments will incorporate findings from our interpretability research, and will validate the effectiveness of ourmethodsusingtesting exercises with intentionally misaligned models. We will publishourfindings in our system cards or Risk Reports. **After:** Read more
#36 (p.11 → p.14)	Replaced (Text)	**Before:** Read more **After:** Policy Target: December 15, 2026
#37 (p.11 → p.14)	Replaced (Text)	**Before:** Cross-cutting Target: January 1 **After:** Policy proposal on embedded third-party evaluators Independent evaluation is a core element of our Responsible Scaling Policy and our policy framework for mitigating catastrophic AI risks. We recently committed to going further by embedding independent evaluators, giving them deeper access to models and internal information than external evaluators typically have, while preserving their independence.
#38 (p.14)	Added (Text)	We will develop and publish a policy proposal on how governments can support deeper access for independent evaluators, including how evaluators are funded, what access they need, and how their independence is protected. We will draw on our own experience working with them.
#39 (p.14)	Added (Text)	Policy Target: January 15
#40 (p.11 → p.14)	Replaced (Text)	**Before:** Moving toward an “eyes on everything”state for our internal AI development activities For many reasons - preventing theft or sabotage of our models, ensuring adherence to Claude’s Constitution, and ensuring that our model training hasn’t been manipulated by insiders or attackers (or AI models themselves) - we want our internal environment to be intensively monitored and comprehensively understood. We aim to establish comprehensive, centralized records of all our critical AI development activities, and to use AI to analyze these records for issues including concerning behavior by insiders (both human and AI) and security threats. **After:** Policy framework for targeting the pace of AI development Our current policy framework focuses on blocking or deterring the deployment of dangerous AI systems. However, we believe policymakers may also need to target the pace ofAI development directly, because the automation ofAI R&D could accelerate progress faster than governance or safety measures can keep up. Speed itself could become an important risk factor.
#41 (p.14)	Added (Text)	We are working on additional policy proposals that provide options for policymakers to address the risks raised by the speed ofAI development, particularly under conditions of recursive self-improvement. We will either publish a proposal or publish an explanation of why we decided not to.
#42 (p.14)	Added (Text)	Alignment
#43 (p.14)	Removed (Text)	Read more

#44 (p.14)	Removed (Text)	Policy
#45 (p.14)	Removed (Text)	A roadmap for policymakers update We recently completed our “roadmap for policymakers” goal and expect to publish a new goal in the coming weeks.
#46 (p.15)	Removed (Text)	
#47 (p.15)	Added (Text)	Update on “Upholding Claude’s Constitution”
#48 (p.15)	Added (Text)	We are retiring this goal in its current form while we spend more time reflecting on which goals are most effective for alignment, and we will publish more about our thinking in the future. At a minimum, we plan to continue to update the public Constitution to match the most recent trained-on Constitution within 90 days of relevant internal or external deployments of models trained on updated Constitutions. See more on our updates page.
#49 (p.16 → p.17)	Replaced (Text)	**Before:** Products **After:** Research
#50 (p.17)	Replaced (Text)	**Before:** Claude **After:** Policy
#51 (p.17)	Replaced (Text)	**Before:** Claude Code **After:** Commitments
#52 (p.17)	Replaced (Text)	**Before:** Claude Code Enterprise **After:** Company
#53 (p.17)	Replaced (Text)	**Before:** Claude Cowork **After:** Products
#54 (p.17)	Removed (Text)	@Claude
#55 (p.17)	Removed (Text)	Claude Design
#56 (p.17)	Removed (Text)	Claude Science
#57 (p.17)	Removed (Text)	Claude Security
#58 (p.17)	Removed (Text)	Claude in Chrome
#59 (p.17)	Removed (Text)	Claude for Microsoft 365
#60 (p.17)	Removed (Text)	Skills
#61 (p.17)	Removed (Text)	Download app
#62 (p.17)	Removed (Text)	Pricing
#63 (p.17)	Removed (Text)	Log in to Claude
#64 (p.17)	Removed (Text)	Mythos
#65 (p.17)	Removed (Text)	Fable
#66 (p.17)	Removed (Text)	Opus
#67 (p.17)	Removed (Text)	Sonnet
#68 (p.17)	Removed (Text)	Haiku

#69 (p.17 → p.18)	Replaced (Text)	**Before:** AI agents **After:** Help and
#70 (p.17)	Removed (Text)	Code modernization
#71 (p.17)	Removed (Text)	Coding
#72 (p.17)	Removed (Text)	Commerce
#73 (p.17)	Removed (Text)	Customer support
#74 (p.17)	Removed (Text)	Cyber
#75 (p.17 → p.18)	Replaced (Text)	**Before:** Enterprise **After:** Terms and policies
#76 (p.17 → p.18)	Replaced (Text)	**Before:** Financial services **After:** Our mission Anthropic is a Public Benefit Corporation. Our mission is to ensure the world safely makes the transitionthroughtransformative AI.
#77 (p.18)	Replaced (Text)	**Before:** Government **After:** Log in Claude.ai Claude console
#78 (p.18)	Replaced (Text)	**Before:** Healthcare **After:** © 2026 Anthropic PBC
#79 (p.18)	Removed (Text)	Higher education
#80 (p.18)	Removed (Text)	K-12teachers
#81 (p.18)	Removed (Text)	Legal
#82 (p.18)	Removed (Text)	Life sciences
#83 (p.18)	Removed (Text)	Nonprofits
#84 (p.18)	Removed (Text)	Sales
#85 (p.18)	Removed (Text)	Small business
#86 (p.18)	Removed (Text)	Claude Platform
#87 (p.18)	Removed (Text)	Overview
#88 (p.18)	Removed (Text)	Developer docs
#89 (p.18)	Removed (Text)	Pricing
#90 (p.18)	Removed (Text)	Ecosystem
#91 (p.18)	Removed (Text)	Marketplace
#92 (p.18)	Removed (Text)	Regional compliance
#93 (p.18)	Removed (Text)	Claude on AWS
#94 (p.18)	Removed (Text)	Google Cloud
#95 (p.18)	Removed (Text)	Microsoft Foundry
#96 (p.18)	Removed (Text)	Console login
#97 (p.18)	Removed (Text)	Resources
#98 (p.18)	Removed (Text)	Blog

#99 (p.18)	Removed (Text)	Claude partner network
#100 (p.18)	Removed (Text)	Community
#101 (p.18)	Removed (Text)	Connectors
#102 (p.18)	Removed (Text)	Courses
#103 (p.18)	Removed (Text)	Customer stories
#104 (p.18)	Removed (Text)	Developer blog
#105 (p.19)	Removed (Text)	Engineering at Anthropic
#106 (p.19)	Removed (Text)	Events
#107 (p.19)	Removed (Text)	Plugins
#108 (p.19)	Removed (Text)	Powered by Claude
#109 (p.19)	Removed (Text)	Service partners
#110 (p.19)	Removed (Text)	Tutorials
#111 (p.19)	Removed (Text)	Use cases
#112 (p.19)	Removed (Text)	Programs
#113 (p.19)	Removed (Text)	Startups
#114 (p.19)	Removed (Text)	Scientists
#115 (p.19)	Removed (Text)	Help and security
#116 (p.19)	Removed (Text)	Availability
#117 (p.19)	Removed (Text)	Status
#118 (p.19)	Removed (Text)	Support center
#119 (p.19)	Removed (Text)	Company
#120 (p.19)	Removed (Text)	Anthropic
#121 (p.19)	Removed (Text)	Careers
#122 (p.19)	Removed (Text)	Leadership
#123 (p.19)	Removed (Text)	Policy
#124 (p.19)	Removed (Text)	Economic Futures
#125 (p.19)	Removed (Text)	Research
#126 (p.19)	Removed (Text)	News
#127 (p.19)	Removed (Text)	Claude’s Constitution
#128 (p.19)	Removed (Text)	Claude Corps
#129 (p.19)	Removed (Text)	Keep thinking
#130 (p.19)	Removed (Text)	Policy on the AI Exponential
#131 (p.19)	Removed (Text)	Responsible Scaling Policy
#132 (p.19)	Removed (Text)	Security and compliance

Generated by Diffchecker

#133 (p.20)	Removed (Text)	Transparency
#134 (p.20)	Removed (Text)	Terms and policies
#135 (p.20)	Removed (Text)	Privacy choices
#136 (p.20)	Removed (Text)	Privacy policy
#137 (p.20)	Removed (Text)	Consumer health data privacy policy
#138 (p.20)	Removed (Text)	Responsible disclosure policy
#139 (p.20)	Removed (Text)	Terms of service: Commercial
#140 (p.20)	Removed (Text)	Terms of service: Consumer
#141 (p.20)	Removed (Text)	Terms of Service: US K-12
#142 (p.20)	Removed (Text)	Data Processing Agreement: US K-12
#143 (p.20)	Removed (Text)	Usage policy
#144 (p.20)	Removed (Text)	©2026 Anthropic PBC. Services in the EU are provided by Anthropic Ireland Limited.