Announcements

# Redeploying Fable 5

**UPDATE**

**Claude Fable 5 and Mythos 5 redeployed**

1 july 2026

Access to Claude Fable 5 and Mythos 5 is now restored.

On Friday, June 12, the US government applied export controls to our newest models, Claude Fable 5 and Claude Mythos 5. This required us to restrict access to foreign nationals, whether inside or outside the United States. Because the order took effect immediately and we had no reliable way to verify nationality in real-time, we suspended access to both models for all users.

**As of today, June 30, the export controls on Fable 5 and Mythos 5 have been lifted.**

Fable 5 will be available starting tomorrow, Wednesday, July 1, to users globally on the Claude Platform, Claude.ai, Claude Code, and Claude Cowork. For Pro, Max, Team, and select Enterprise plans,[1] Fable 5 will be included for up to 50% of weekly usage limits through July 7, after which it will be available via usage credits. We will re-enable access on AWS, Google Cloud, and Microsoft Foundry as quickly as possible.

We have also restored access to Mythos 5 for a set of US organizations, following the US government's approval on June 26. We continue to coordinate with the government to expand access to the broader set of domestic and international partners in the Glasswing program.

In the remainder of this post, we provide further details and updates in four areas:

1. *A timeline of events, including updates we made to our safeguards*. We discuss the events that led to the export control directive and how we addressed it with new safeguards.
2. *Our general approach to safeguards*. We provide more context on how we use safety classifiers to detect potentially dangerous cybersecurity uses of our models.

3. *A shared industry framework*. Although we have reached a constructive resolution, these events have made clear that the industry needs a consistent way to assess and fix potential "jailbreaks" of AI models (techniques that bypass a model's safeguards).[2] A shared standard for judging the severity of a given jailbreak would help AI developers triage new findings as they arise, launch highly capable models with greater safety, and communicate the level of risk consistently to government and industry partners. Together with Amazon, Microsoft, Google, and other Glasswing partners, we've started to develop such a framework, and we outline it below.
4. *Deeper government collaboration*. We're also strengthening our level of collaboration with the US government on new pre-release testing, information sharing, and research collaboration. We describe this deeper collaboration in the final section.

## Timeline and safeguard updates

We released Fable 5 and Mythos 5 on Tuesday, June 9. They both share the same underlying model, but Fable 5 was released with strong safeguards to make it safer for general use. Mythos 5, which has fewer safeguards, was only released to a small number of trusted Project Glasswing partners for use in defensive cybersecurity.

The export control directive on June 12 came after the government became aware of a report in which Amazon researchers had found a method of bypassing Fable 5's safeguards: prompting it so that it identified a number of software vulnerabilities. In one case, the model produced code demonstrating how the relevant vulnerability could be exploited. Over the past two weeks, we have worked closely with the government and other partners, including Amazon, to review the report and evidence.

Our testing confirmed that many less capable models—including Claude Opus 4.8, GPT-5.5, and Kimi K2.7—could identify the same vulnerabilities as Fable 5 did in the report. When it came to the demonstration of how to exploit the single vulnerability, every model we tested could produce the same demonstration as Fable 5 (including Claude Haiku 4.5, Sonnet 4.6, Opus 4.6, Opus 4.7, Opus 4.8, GPT-5.4, GPT-5.5, and Kimi K2.7).

Importantly, the reported technique did not expose any unique Mythos-level cyber capabilities. The behavior reflected a borderline case for Fable 5's safeguards—as we will explain below, there are some tasks that are unlikely to be dangerous but are nonetheless blocked by the safeguards out of an abundance of caution. The reported technique allowed access to one such behavior, but it only involved routine defensive cybersecurity work.

Even so, we moved quickly to address the reported bypass. Working closely with the government, we trained an improved safety classifier that targets and blocks the behavior described in the report. Users will be notified if a request to Fable 5 is blocked, and the request will instead be sent to Opus 4.8.

The new classifier means that the specific technique described in the Amazon report is blocked in over 99% of cases. In a very small fraction of cases the model may provide information that isn't detailed enough to help a cyberattacker. As we describe below, the model's safeguards are not expected to block *all* low-risk routine cyberdefense capabilities—just those that are potentially harmful. Researchers from the US Department of Commerce's Center for AI Standards and Innovation (CAISI) have tested both our prior and new safeguards and agree that they are extraordinarily strong.

The new classifier also comes at the cost of flagging benign requests more often during routine coding and debugging tasks. As with all our safeguards, we'll continue to refine this to better distinguish genuine misuse from legitimate requests and reduce false positives.

## Our approach to cybersecurity safeguards

Claude Mythos 5 can be used to find and exploit software vulnerabilities more effectively than any other model—and all but the most skilled human security experts. These prodigious cybersecurity capabilities make it uniquely attractive to malicious actors who wish to misuse it in cyberattacks.

Claude Fable 5, however, provides no such unique offensive capabilities. This is because we launched it with the strongest safeguards we've ever applied to a model. In the month prior to launch, we transferred staff from various teams within Anthropic to double the number of researchers and engineers working on this problem.

Fable 5 launched with a variety of safety mechanisms, each of which alone does not provide perfect defense but when combined make the model very difficult to misuse (an approach known as "defense in depth"). Some defenses involve training the model to decline to assist with dangerous requests; others involve retroactively analyzing patterns of misuse.

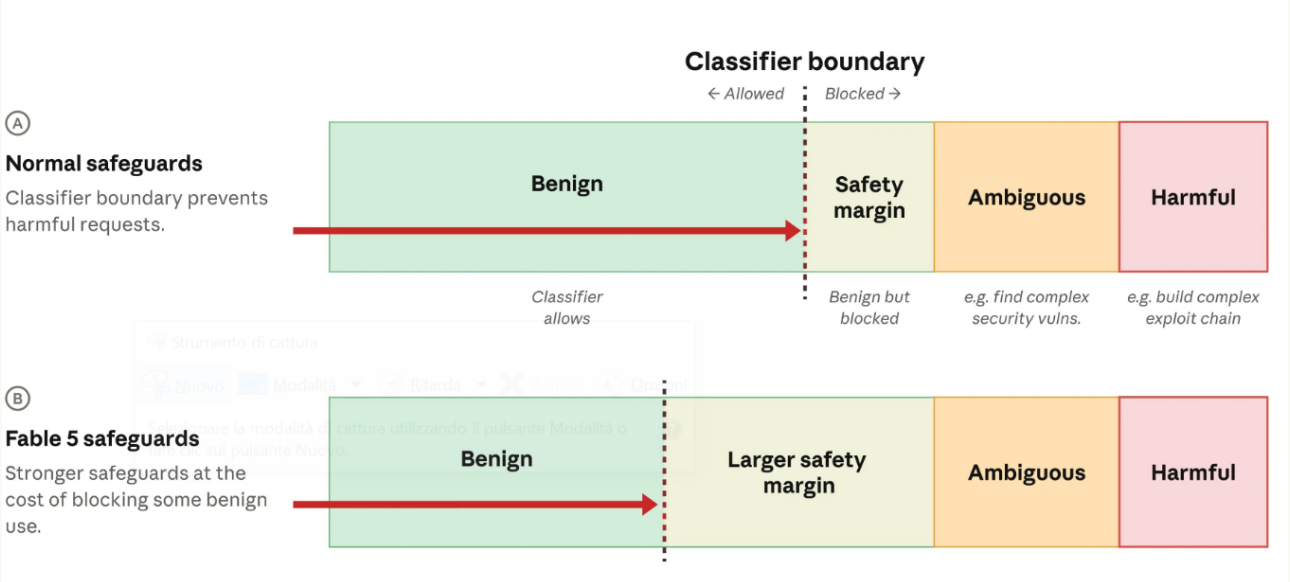
One particularly important safety mechanism involves *classifiers*—smaller automated AI systems that, during an interaction, detect when the model is asked to perform a potentially

harmful cybersecurity task (or produces potentially harmful outputs). When this occurs, the classifiers block the model from responding to requests. The ultimate goal of these classifiers is to prevent the model from engaging in uniquely dangerous behaviors.

Like all safety mechanisms, classifiers can make mistakes. They sometimes fail to notice potentially dangerous content, and in some cases they can be deliberately "jailbroken": users can prompt the model in unusual ways to trick the classifiers and get the model to produce harmful outputs that the system should have blocked.

We therefore deliberately set the safety classifiers to trigger on a set of requests that we know are likely benign. This "safety margin" approach means that a request has to look very clearly safe to avoid triggering the classifier (see row A in the diagram below). Users experience the safety margin as a model refusing to respond to some reasonable, non-harmful requests.

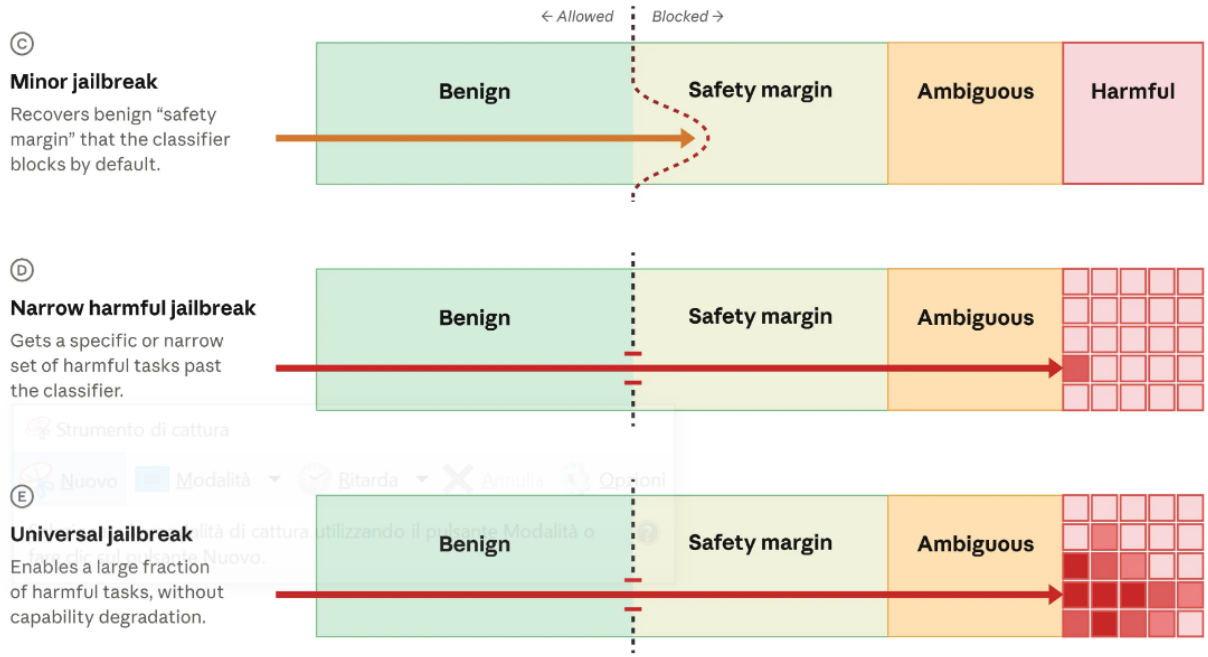
For Fable 5, we made this safety margin much larger than in any prior launch (row B), meaning that many more benign requests would be blocked. We understood that these kinds of false positives would be frustrating for users, but made this tradeoff in the interest of making the model's other capabilities widely available.




**An illustration of our cybersecurity safety classifiers.**
When a request is made to the model, the classifiers detect whether it is benign (and allowed), or potentially harmful (and blocked). The classifiers block ambiguous requests (those that are clearly to do with cybersecurity but could potentially be for defensive purposes, like finding security vulnerabilities) and harmful requests (those that are clearly dangerous, such as a request to build a chain of software exploits). As shown in row A, we also include a "safety margin", where the classifier will block requests that are probably benign but have some small chance of being harmful. This increases our confidence that all harmful requests will be blocked. For Fable 5 (row B) we made the safety margin even larger, meaning that more benign requests would be blocked—but fewer genuinely harmful requests would be missed. "Vulns" = vulnerabilities.

The safety margin also helps mitigate jailbreaks. Many jailbreaks are narrow: they unblock a very specific model behavior but nothing more. In some cases, a hypothetical user can jailbreak the model in a minor way and intrude into the safety margin (or sometimes into ambiguously harmful behavior), but not to the core harmful behaviors that we aim to block (row C below). Our view is that jailbreaks of Fable 5 reported so far fit into this minor category.

More serious jailbreaks unblock more harmful behaviors. Narrow harmful jailbreaks (row D) can elicit some specific harmful behaviors. These jailbreaks are typically of low to moderate severity, because the narrowness limits the attacker. The most concerning category is a *universal* jailbreak (row E), which unblocks a wide range of harmful behaviors.




**How jailbreaks interact with our safety classifiers.**
In the case of a minor jailbreak (row C), the classifiers do not block the request, but the request is still within our safety margin (and is thus very unlikely to be harmful). In a narrow harmful jailbreak (row D), the prompt breaches the classifiers and unblocks a specific harmful behavior from the model. In a universal jailbreak (row E), a prompt unblocks an entire class of harmful behaviors.

As we noted when we launched Fable 5, it is probably impossible to make any AI model fully robust (that is, impervious) to jailbreaks.[3] We expect that some jailbreaks will be found for our models, and that they will vary in severity: there will be many minor jailbreaks,

some narrow harmful ones, and although no universal jailbreaks for Fable 5 have been discovered at the time of writing, expert safety researchers continue to red-team it. We seek to ensure that we and our safety partners will be the first to find major jailbreaks and fix them before malicious actors can use them for harm.

The cautious approach outlined above means that the vast majority of jailbreaks will not successfully unblock dangerous behaviors. Our classifiers make successful jailbreaks very costly and high-effort to produce, and even *if* a jailbreak is successful, our extra layers of defense provide additional mitigation. We'll continue to update our classifiers as we learn more about novel jailbreak techniques.

## A consensus industry framework for jailbreaks

There's currently no consensus in the AI industry on how to describe, in objective terms, the severity of an AI jailbreak. This adds a great deal of uncertainty whenever a new jailbreak technique is discovered: developers have no agreed-upon standard for which findings to focus on most urgently, and governments have no agreed-upon standard for when to act.[4]

This problem will become more acute in the coming months, as more models with powerful cybersecurity (and other) capabilities are trained, assessed, and released. A common standard for assessing AI jailbreaks would help us and other companies launch new models safely, as well as allow our users to make the most of their advanced capabilities.

We are therefore partnering with Amazon, Microsoft, Google, and other Glasswing partners to draft a consensus framework for assessing the severity of AI jailbreaks and how AI developers should respond to them. We invite other industry partners and model providers to join us in this effort.

Our current proposal is to score a given jailbreak on the four different criteria below. The first two describe what the jailbreak provides to the attacker; the latter two describe how quickly the jailbreak can become a real-world problem:

1. *Capability gain*. How far beyond existing tools does the jailbreak take the user? If existing widely available tools (including other, weaker AI models) can reach the same capability as the jailbroken model, the score here will be low; if the jailbreak unblocks model capabilities that can significantly accelerate even domain experts, the score will be high.

2. *Breadth of capability gain*. For how many distinct offensive tasks does the same jailbreak technique work? Cases where the jailbreak only allows the model to pursue narrow targets will score low; cases where the same jailbreak technique works for multiple different targets or techniques will score high.
3. *Ease of weaponization*. How much human effort does it take to turn the jailbreak into an attack? Where the jailbreak involves a great deal of skilled prompting and many retries, the score will be low; where the jailbreak works on a single prompt or on the first or second try, the score will be high.
4. *Discoverability*. How easy is it for someone to obtain the technique? If it requires specialist knowledge it will score low; if it is already widely known and available online it will score high.

We propose to use this severity framework to calibrate our response to newly discovered jailbreaks. For the most severe class of jailbreaks (e.g., a jailbreak that, among other characteristics, is being used to actively cause a devastating impact on critical power grids or banking systems), we will immediately begin deploying preliminary mitigations upon confirmation of severity. We are also creating a team to provide 24/7 monitoring of key jailbreak submission channels.

Any method of scoring jailbreaks will be imperfect. Still, there is value in being able to communicate the approximate severity of a given finding through a common framework. This is a work in progress; as we receive feedback from more partners, we expect the framework to evolve over time.

We expect to share more details on the proposed framework soon. In the meantime, we're also launching a new HackerOne program where security researchers can submit potential cyber jailbreaks they've discovered in Fable 5 (once available) for our review.

## Partnering with the US government on frontier AI security

Over the past ten weeks, Anthropic has worked closely with the US government as it developed the approach reflected in the June 2 Executive Order on *Promoting Advanced Artificial Intelligence Innovation and Security*. Our engagement spanned the Office of the National Cyber Director, the Office of Science and Technology Policy, the Department of the Treasury, the Department of Commerce (including CAISI), and relevant national security agencies.

We are committed to continuing that work, building on nearly two years of pre-existing collaborations with US government partners on pre-deployment testing and evaluation. The commitments below reflect both that pre-existing work and our new proposals to scale up our government collaboration as the above framework is finalized:

1. *Pre-release government access and evaluation.* For models that materially advance the capability frontier in areas relevant to national security, we will provide designated government partners with expanded early access to both the models and the safeguards that accompany them. Those partners can then run independent capability evaluations and test our guardrails before broad release. We will dedicate Anthropic technical staff to work alongside government evaluators during these testing periods.
2. *Rapid information sharing on safeguards.* When significant jailbreaks or misuse patterns are identified, we will quickly investigate, triage, and notify appropriate government counterparts. We will share the new safeguards we build in response so they can be independently tested. We will also provide government partners with our threat intelligence reporting in advance of publication and participate in the interagency cybersecurity vulnerability clearinghouse established under Sec. 2(d) of the June 2 Executive Order.
3. *Dedicated resources for joint research.* We are substantially scaling up joint work with government partners on AI security. We will stand up dedicated Anthropic teams to work on shared government priorities, provide a significant compute allocation to support government testing and research, and make our safety and red-teaming expertise available to help advance the state of the art in AI evaluation.
4. *A common industry bar.* We will work with the government and with industry peers toward a shared, voluntary security and evaluation standard for frontier model providers. We'll contribute evaluations, tooling, and best practices that the government can apply across the field.

Our hope is that this collaboration, along with our proposed consensus industry framework, will serve as the basis for systematic rules for the whole industry—and even offer the beginnings of a template for effective global coordination on the risks and benefits of AI.

These rules should be codified in strong regulation and applied equally across frontier model developers. Government involvement in AI releases requires a durable, transparent process that gives cyber defenders and others the certainty they need about access to powerful models.

We look forward to deepening our government collaboration in the ways we've described above. We're also grateful to our users for bearing with us through this disruption, and to the researchers and industry partners who worked alongside us to make Fable 5 and Mythos 5 available again.

## Footnotes

1. For standard Enterprise seats, there is no included Fable 5 allowance, although you can get access through usage credits. If credits are not enabled, your users will not have access to Fable 5. For premium Enterprise seats, through July 7, Fable 5 is included in your subscription. It draws from each member's seat usage at no additional cost. After July 7, your team can continue using Fable 5 by enabling usage credits. If credits are not enabled, your users will no longer have access to Fable 5.
2. Note that sometimes the term "bypass" is itself used instead of "jailbreak." For current purposes, we consider these to be synonyms, but for the remainder of this article we use "jailbreak" because (a) this is a more commonly used term and (b) it is consistent with the terminology we have used in previous work.
3. Analogously, no piece of software is immune to vulnerabilities (though in general, software vulnerabilities are more straightforwardly discovered and patched than LLM jailbreaks).
4. In other areas of security research, there *are* agreed-upon standards: for example, the Common Vulnerability Scoring System (CVSS) is a common way of assessing the severity of a given software vulnerability.