

Anthropic reports Claude submitted forms and bypassed limits on government sites

Source copy kept by wheresthe.ai for the story “Anthropic reports Claude submitted forms and bypassed limits on government sites” (<https://wheresthe.ai/d/anthropic-unintended-model-actions-2026-10/>).

Source: <https://www.anthropic.com/research/investigating-unintended-model-actions>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 704b9f7389f02dc34fc58ecbf7d0a8c5328def54b068eefb7452fdf81c3272

Documents the source offers: <https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf>

Alignment Investigating unintended model actions in our evaluations and internal use Oct 9, 2026

This report describes examples of unintended model actions we’ve observed during evaluations and internal use of Claude. It is part of our effort to publish more frequent standalone reports on model behavior and alignment beyond our system cards, which we publish with each model release, and our risk reports, which we publish every three to six months as part of our Responsible Scaling Policy. We believe it’s important to be transparent about what we see our models do during testing and use. The behaviors can be grouped into four categories: Claude exploiting a basic flaw in software to run commands on a server; Claude submitting a sensitive form on a real website when it should not have; Claude working around a restriction to reach data that was gated by a token or a fee; and Claude using URL shortening services to get around limits in its fetch tool. We have chosen not to name the organizations involved in the examples below to avoid exposing vulnerabilities in their systems, and at their request. For this reason, we also provide less detail about each case than we otherwise would. Some of the cases described below involved websites run by U.S. government agencies at the federal, state, and local levels. We have briefed the White House on these cases and notified each agency involved. The cases we’ve identified to date in these categories had minimal real-world impact. We consider these behaviors to be significantly less severe from an alignment and security perspective than the cybersecurity incidents we reported on July 30 and September 9 . They resemble behaviors that we’ve described in our system cards since Claude Mythos Preview . Most are forms of persistence , in which Claude, when it cannot complete a task as given, works around a restriction instead of stopping. Although the impact of these behaviors was minimal and we had already turned off live internet access for some high-risk and cybersecurity evaluations, we have now decided to expand that to include all our internal evaluations until we have confirmed that our security and monitoring measures (described in the remediation section of this post) reliably catch behaviors like these. Below, we also discuss alignment considerations and give more detail on how we’re modifying training to reduce the likelihood of further misbehavior. We identified most of these cases through a review of transcripts that we began in July. Our review first focused on our cybersecurity evaluations—tests where a model is deliberately asked to probe or attack a test system, and where internet access is meant to be disabled. We have since extended our scanning to encompass a much wider range of instances where Claude could have reached the internet, including tests where internet access is deliberately enabled so Claude can be evaluated on real-world tasks. We began by looking for incidents of similar severity to the cybersecurity incidents we reported this summer; we have not found any to date. We then broadened the search to lower-severity cases, where a model interacted with real websites or systems in ways we didn’t intend. We are also now scanning a much larger pool of lower-risk transcripts, as well as our use of Claude within Anthropic and in reinforcement learning (RL) environments where Claude has access to the internet. As this work continues, we plan to report new instances of unintended behaviors. All cases reported here involved Claude interacting with the outside world; to our knowledge, none of them involved customer data or Anthropic’s own internal systems. In this post, we explain why we run evaluations, which are where most of these cases occurred; describe each behavior in more detail; and offer a preliminary view of what the behaviors suggest about Claude’s alignment and how we’re mitigating them. Why and how we run evaluations The behaviors described in this post are not specific to evaluations, but many of the cases we’ve identified to date occurred during evaluation runs. We continually test models on a wide range of tasks prior to releasing them. We use many different evaluations (standardized sets of tasks, scored the same way each time), each of which helps to paint a picture of Claude’s skills in a particular area. Many of the evaluations we use are public; they are written by outside researchers and can be run by any developer, allowing us to compare capabilities across models. Other evaluations we build in-house. Because language models are non-deterministic—that is, their responses always involve some element of randomness, and they may carry out the same task slightly differently each time—we have Claude complete each evaluation task hundreds or thousands of times (each attempt is called a run). Testing so many times allows us to

understand how a model typically performs, and also to catch rare cases where a model does something we don't expect. What we learn from evaluations shapes decisions across Anthropic, including how we train models, what safeguards we build around them, and when we release them. Evaluations are also where a model's behavioral issues are sometimes first revealed. Models learn much of what they can do through RL, in which they attempt a task within a specific environment many times over and are rewarded when they succeed. However, training environments are imperfect. If training rewards something we didn't intend—such as finding loopholes or working around a restriction—the model learns that the workaround pays off and may then apply it elsewhere. This is known as reward hacking. While we have processes in place to identify and filter out reward hacking during training, evaluations are a critical mechanism for understanding a model's behavioral propensities in different contexts and on different sets of tasks than those we have in training. Some tasks—such as searching the web for hard-to-find information—are difficult to realistically simulate in an environment without internet access. Running them with access has been standard practice within the industry: public web search benchmarks are run on the live internet by default, and running them the same way allows us to compare our models with other models. Most of the cases in this report occurred during such evaluations.

Overview of behaviors Below, we describe four categories of behavior in which Claude acted on real websites and systems in ways we did not intend, each of which involved an organization or person outside Anthropic.

Exploiting a basic flaw in software to run commands on a server When Claude could not complete an evaluation task directly, either because its own tools were deliberately limited or because an outside service it needed was unavailable, it used other tools hosted on a third party's site to achieve its goal. In some circumstances, it did this by exploiting basic flaws in that site's software, using SQL or command injection (a way of sending input that the site mistakenly runs as instructions). For example, one evaluation asked Claude Mythos Preview to run a scientific analysis. The public tool it needed to perform that analysis was hosted by a university; when the model sought to access it, the tool returned an error. Looking for another way to complete the task, Claude explored the tool's website, found a script on the university's server that would return any file it was asked for, and used it to copy files from the server, including the script's own code. In that code Claude found the injection flaw that let it run commands on the server, and used it to run the calculation. In another instance, the data Claude Mythos 5 needed was free, but Claude could not access it without accepting a data use agreement, which it did not have the tools to do. Claude used applications hosted on the website to load the page and accept the agreement on its behalf. We saw this behavior on DeepSearchQA and BrowseComp (both public web research evaluations), on LABBench2 (a public biology research benchmark), and during internal evaluations. Submitting a form it should not have We saw instances where Claude submitted an online form when it shouldn't have. Generally, this occurred when an evaluation's instructions were ambiguous, or when a misconfiguration within the environment prevented Claude from working with dummy forms. For example, in a case that happened multiple times on the same evaluation, an unreleased, non-frontier research model was meant to fill out a practice copy of a government form. When the copy failed to load or the model closed it by mistake, the model instead navigated to the website where the real form is normally hosted and submitted the form there. In another evaluation, Claude Haiku 4.5 was instructed to complete a form up until the final submission stage, but to stop before actually submitting. Several times Claude mistakenly submitted the form instead, expecting there to be an additional confirmation page. In these instances, the model's chain-of-thought stated that it was demonstrating the process, not submitting a real request. In a third example of this behavior, Claude Haiku 4.5 had been tasked with generating and performing example tasks on randomly selected webpages. In one run, the model landed on a page referencing an unsolved homicide; that page contained a tip form run by a police department. Claude was instructed never to log in, create accounts, enter personal data, make purchases, or submit anything destructive, but the instructions did not rule out form submissions. Claude filled out the form with the following: "I may have information regarding this case. I recall seeing someone matching the description in the area around [the street named on the page] during that time period. Please contact me if this information is relevant." (The website did not include a description of the perpetrator.) The model left the name and contact fields empty, which the form allowed, and submitted it. The submission was flagged as spam and was never forwarded for investigation. We saw this behavior on OSWorld (a public computer use evaluation), on Odysseys (a long-horizon task evaluation), and during internal usage.

Working around restrictions to reach gated data Claude sometimes worked around a restriction to reach data to which it did not have access. Generally, this occurred when a server refused Claude's request, or when the data was offered only for a fee. For example, in an evaluation that asked Claude Mythos 5 to identify a location shown in a photo, it attempted to access a local government's property map to triangulate its guesses. Claude could read webpage text, but the evaluation limited its ability to click through it as a person would. Instead, it read the settings file the site sends to browsers, found working access tokens, and used them to send requests directly to the server behind the map to access the data. In another example of this behavior, an Anthropic researcher used Claude Mythos 5 on a

statistics project that required it to pull public data that was available from a state agency for a fee. Instead, Claude learned from an archived copy of the agency's website that its public dashboard issues an access token to any visitor. It requested one and used it to query the database without paying the fee. In both examples, Claude obtained data already available to the public but only for a fee. We saw this behavior on Humanity's Last Exam, an evaluation of expert-level academic questions, and during internal usage. Using URL shortening services

Some of our fetch tools, which let Claude read webpages, limit the length of the URLs Claude can request. This limit is in place to reduce the risk from Claude using long URLs to take certain unwanted actions, such as exfiltrating larger amounts of data. We saw several models, including Claude Opus 5 and Claude Mythos 5, get around this limitation by using free URL shortening services. We first identified this behavior internally, but as we were writing this post, an operator of the da.gd URL shortening service let us know that they had also found Claude using their website for this purpose. We thank the operator for sharing these details. Remediation We've taken several preventive measures. Some of the public evaluations we no longer run; others we have moved to their offline versions, or rebuilt them so that their tasks do not reach live websites. We've also made broader changes. We have updated the guardrails on some of our internet access tools, such as the web fetch tool, to heavily restrict what the model can do with them. And, as part of our ongoing agent security and monitoring efforts, we have built tooling to automatically detect and block the kinds of behaviors described above. This tooling now runs on most of our evaluations and on internal agentic use of frontier models. When we tested it against the cases described in this post, it blocked all of them. As we described in our August 31 post, we are also continuing to fix or remove training environments that reward Claude for working around tool restrictions or other blockers, so that they do not incentivize these behaviors or permit reward hacking. We're also taking other steps to prevent unintended agentic actions within Anthropic, and to catch them if they occur. These include migrating internal agents to centrally managed infrastructure with strong containment, minimizing internet access for internal agents and training processes, and monitoring far more of what agents do through techniques like safety classifiers and hierarchical summarization. These measures are now part of our security team's detection and response procedures so that we can respond to and contain undesired behavior quickly. We expect to publish more on these approaches, and to build them directly into our products.

Discussion and alignment takeaways While we have not completed a full alignment assessment of these cases, we consider them to be less severe than the cybersecurity incidents from this summer. We explain why below, and share the general lessons we've drawn from them.

Alignment context As we establish a regular public reporting process for behaviors like these, we also want to situate them in terms of their severity. This remains challenging: judging the extent of an alignment failure often depends on what the model was trying to do, and a model's own account of its reasoning is not necessarily reliable evidence of its beliefs or reasons for action, as we discussed in our September 9 alignment assessment. As models grow more capable, they can interact with the world in more sophisticated ways, which makes more complex failures possible. For example, the biased reasoning we described in that assessment was not new, but it led to more serious consequences in a setting where models could capably act on real systems for hours at a time. An alignment severity framework built around today's failures could become outdated quickly, given how fast AI capabilities are advancing. Nonetheless, we found two alignment dimensions useful for examining the cases in this post: overreach, or how far the model went beyond what its task intended, and dishonesty, or whether it gave a misleading account of its actions or intentions. On overreach, we consider these cases to be substantially less concerning than the incidents we reported this summer, in which Claude gained access to real third-party systems for hours during cybersecurity evaluations. In the cases we report here, in contrast, Claude bypassed access controls to reach gated but publicly available data, or exploited software flaws and ran commands on servers (mainly to reach non-sensitive data). On dishonesty, the comparison is more mixed. In one case, tasked with generating example interactions with websites, Claude submitted an invented tip through a police department's online form. From the transcript, Claude appears to have only been producing example content for the task, rather than trying to mislead anyone to achieve a goal. In the most serious incident from this summer, by contrast, Claude's misleading reasoning was sustained over hours and supported its continued attack. Still, judging dishonesty with confidence typically requires a deeper assessment than we have done here, such as replaying transcripts with modifications to test what drove the model's behavior, so our view of these cases may change with further analysis.

Lessons and next steps None of the behaviors we've described here are new, and they do not change our overall view of Claude's alignment. In many of the cases where we observed these behaviors, Claude had been given tasks that were ambiguous or impossible to complete. We and others have observed that when models are given impossible-to-complete tasks, they will pursue unintended and sometimes misaligned strategies to achieve their goals. It is possible that some of these failures could have been avoided if the evaluation questions had more clearly stated what was in and out of scope for the exercise, including the targets, permitted actions, and network boundaries (i.e., what the model should and shouldn't access). However, Claude

encounters ambiguous and impossible tasks every day in real use, and, indeed, several of the cases we observed occurred during regular agentic use of Claude. Behavioral and alignment training is the main technique we have at our disposal to improve Claude's judgment and ability to handle ambiguous situations with care. Historically, we have been more focused on teaching models to respect boundaries and be appropriately cautious in coding environments. We are now expanding these environments to applications like search and computer use, which were involved in the cases described here. However, alignment training is not yet sufficient or fully robust on its own, at least in the short term, so we also rely on defense-in-depth approaches, including the classifiers and safeguards described above. We plan to keep reporting concerning behaviors as our scan and analysis continue. We hope these reports help other developers check for similar behaviors in their own models, since many of the evaluations involved are public and widely used. While these cases had minimal impact, we do not want to diminish the findings, because the same behaviors could do far more harm as models become more powerful. The larger the role models play in society, the more the public deserves to know how they behave. Note: The tip form example described above involved the Philadelphia Police Department, who self-disclosed today via their press release. We shared this finding with the department on October 8 as soon as our technical review was complete. Corrected Oct 12, 2026: A previous version of this post incorrectly stated that Claude's web fetch limits were put in place to prevent SQL or command injections. In a small number of evaluation runs, a model did use URL shortening services while it carried out an SQL injection or another unintended action, but the limits were not put in place to prevent such attacks.

Related content

[Designing more efficient quantum circuits with Claude](#) We gave Claude plain-language specifications for eight quantum circuits, along with an automated evaluator to check and score its designs. Working largely on its own, Claude improved on the best published circuits, in some cases by an order of magnitude or more, which could bring practical quantum computers closer than previously anticipated. [Read more](#) [Claude reveals ECDSA curve origins](#) The decades-old mystery of the origins of one of the most popular cryptographic standards has been explained. [Read more](#) [The AI expectations gap](#) How will AI reshape the economy in the years ahead? What is the core source of disagreement about AI's future impact on the economy? This note presents new survey evidence on these questions. [Read more](#)