ANTHROP\C

September 15, 2026

The Honorable Lisa Blunt Rochester
United States Senate
Washington, DC 20510

RE: August 6, 2026 Letter Regarding Claude Model Behavior in Cybersecurity Evaluations

Dear Senator Blunt Rochester:

Thank you for your August 6, 2026 letter regarding the cybersecurity evaluation incidents we publicly disclosed on July 30, 2026, and the incident the UK AI Security Institute ("UK AISI") reported on August 4, 2026. Anthropic takes these incidents seriously. Since your letter, we have published two further accounts of the July 30 incidents, which we draw on throughout this response. On August 31 we published the changes we have made to our containment, monitoring, and alignment practices. On September 9 we published a detailed alignment assessment of those incidents and of a fourth incident we identified in August, and we released the redacted transcript of the incident involving Claude Mythos 5. (GitHub and PDF versions). We remain committed to a broader industry conversation about how to evaluate increasingly powerful AI models securely, and to sharing what we learn as our review continues.

**The Four Incidents**

On July 21, 2026, OpenAI disclosed that several of its models had escaped an isolated test environment. Beginning July 23, we reviewed past evaluation transcripts for evidence that a Claude model had reached the internet from within an evaluation. We suspended all internal cybersecurity evaluations that same day, as well as external cyber evaluations related to pre-release models, after identifying transcripts indicating that Claude may have accessed the internet while undergoing a cybersecurity evaluation conducted by our evaluation partner, Irregular. After reviewing 141,006 evaluation runs, we confirmed on July 24 three incidents, comprising six evaluation runs, in which a Claude model had accessed the internet and gained access to the production infrastructure of three organizations. We notified our evaluation partner and each affected organization on July 27, and provided the two affected organizations that responded to our notification with the information necessary for them to assess the impact on their systems. On July 30, we notified the security team at PyPI, the public software registry involved in one incident, and voluntarily notified relevant government authorities. We published a detailed public account the same day.

Given the volume of transcripts and our desire to disclose incidents quickly, our initial scan relied on an agentic search. That scan missed a set of cybersecurity evaluation transcripts that also turned out to have internet access. We identified these in August while assembling transcripts to share with METR, an independent AI evaluation organization, and found among them a fourth incident, from January 2026, involving an early version of Claude Opus 4.6. We notified the affected organization, provided it with the details it needs to support its own investigation, notified relevant government authorities, and disclosed the incident publicly on September 9.

The evaluations at issue were capture-the-flag (CTF) exercises run in an environment built and operated by Irregular. The environment was designed as a private, self-contained network of simulated hosts — with no intended connection to the internet — in which a model is given a workstation and instructed to compromise another host on the same simulated network. In each case, the model's prompt stated that it had no internet access, and none stated which systems were in scope or limited where the model could search for the flag. A misconfiguration in the environment nonetheless left the models with a live path to the internet. As is standard for cybersecurity evaluations, the models ran without the cyber safeguards that ship with our publicly deployed models. Neither Anthropic nor Irregular was aware of the misconfiguration until our July review. While we still believe that these incidents would not have occurred had the environments been isolated from the internet as intended, as our September 9 alignment assessment describes, we found that the models' behavior and actions upon reaching the internet represented alignment issues.

The earliest incident occurred in January 2026 and involved an early checkpoint of Claude Opus 4.6, which was given a CTF task in a different challenge within the same third-party evaluation environment as the prior three incidents. The model accidentally broke its target by assigning it a conflicting IP address with another machine, thus making the target unreachable and the task impossible to solve. Recognizing this, the model attempted to abort the task using a command but was unsuccessful due to a misconfiguration in our evaluation harness. The model attempted to give up seven more times across the transcript. Unable to abort the task, the model initially attempted to reach the target via other in-scope means but failed to do so. It then began exploring its environment and found the same egress path it found in incident 3. The model discovered a machine belonging to a third party that it was able to access, and stated that it believed this third party was part of the CTF. Inside the machine, the model found a file listing a password, which it used to gain admin access to the system. The model then harvested further credentials, modified the system's settings to make it easier to access, and read the personal information of one person associated with the third party. The session ended when the model exhausted its token budget. We notified the affected party after we discovered this fourth incident.

The second incident occurred in April 2026 and involved Claude Opus 4.7, across four runs

of the same evaluation. The fictional target company shared a name with a real, active website domain. When the model could not reach its intended target, it found the real company's systems reachable on the internet. It obtained application and infrastructure credentials and accessed a database containing several hundred rows of production data. It then downloaded and modified user records. The model recognized in its reasoning that the company was real, but assumed that it was part of the exercise and almost never questioned or tested that assumption. It stopped when the target's backend stopped responding.

The third incident occurred in June 2026 and involved an internal, general-purpose research model that is roughly similar to Claude Mythos 5 in capabilities. Midway through the task, the evaluation environment automatically shut down the model's target machine, which was configured to run for only 24 hours. The model then searched the open internet for alternatives, scanning roughly 9,000 targets. It concluded, incorrectly, that it was operating on a simulated replica of the internet. It then attempted to access several neighboring systems and, using access tokens and passwords, gained access to one, where it downloaded files and inserted a remote-control script. When it recognized that this system belonged to a real company, it declared the system out of scope and stopped.

The fourth incident occurred in July 2026 and involved Claude Mythos 5. The model registered an email account and a PyPI account and published three versions of a malicious software package to PyPI, the public registry from which most Python software is installed. The package was installed on 15 third-party systems. We believe all 15 belonged to security vendors that routinely install new packages in sandboxed environments to scan them for security problems. One vendor's scanner leaked its access credentials to the model, which then used them to access that vendor's live database. PyPI's own security systems removed the package within an hour. The session ended when it timed out while the model was preparing an additional malicious package.

In none of the four incidents were our own production systems or customer data accessed or affected. The evaluations in all four incidents ran on infrastructure with no access to either.

Beyond these four cases, we are not aware of other cybersecurity evaluation incidents where a model gained unauthorized access to third-party systems. As we shared previously, we're working with METR to conduct a third-party review of these incidents.

**Why This Happened**

In our July 30 post, we described these incidents as closer to a harness and operational failure than a model alignment failure, and said the models appeared to have attacked real targets because they believed those targets were part of the simulation. We continue to believe these incidents would not have occurred had the environments been isolated from the internet as intended. However, as explained in our September 9 alignment assessment,

our investigation identified two alignment issues, present at varying degrees of severity across the four incidents: biased reasoning, in which the models tended to disregard or misinterpret evidence that it was operating on the real internet, and recklessness, or a willingness to take harmful actions in the narrow pursuit of their assigned task. We've described milder forms of these behaviors in our prior system cards, but we consider these cases to be more serious instances than the behavior reported in the system cards. While the incidents represent more serious instances of both forms of misalignment, they do not represent categorically new failure modes of alignment in our models. Our assessment of these incidents did not rely on the models' stated reasoning alone. We also asked the models follow-up questions at points in the transcripts. We re-ran the models from specific points after editing the evidence in front of them ("resampling") and used tools to examine the models' internal states.

Claude Mythos 5 repeatedly said in its reasoning that its environment was simulated. When separate instances of the same model were shown each tool output from the incident on its own, they judged 79% of the outputs to be likely evidence of a real system. In the incident itself, the model clearly labeled evidence as a sign of the real internet in only about 1% of its actions. Adding stronger signs that the environment was real made the model more likely to acknowledge possible real-world harm, but did not significantly change its actions. Only when told unambiguously that it was on the public internet did it abandon its original route to uploading the package. Interpretability analyses weakly suggest that its outward statements did not fully reflect its internal state. We are most concerned by this incident.

The internal research model spent significant time testing whether the internet it could reach was real, but read the evidence in a biased way. Separate instances judged 66% of that evidence to point to the real internet; the model itself labeled 31% that way. Its decision to stop once it recognized a real company appears to have been an unusually favorable outcome: in resampling, it de-escalated from that point in only 5.5% of runs. Unlike Mythos 5, it acknowledged its biased reasoning when questioned.

Claude Opus 4.7 recognized in its reasoning that its target was a real company, but assumed that the company was in scope for the task because it was reachable from the exercise network and shared a name with the fictional target. Its behavior depended mainly on whether it believed it was authorized: when we added clear signs that access was not authorized, it mostly stopped engaging with the real website. We did not investigate the January incident at the same depth, because we identified it recently and it involved an early checkpoint of an older model. From a preliminary assessment, we do not consider it more severe than the other three.

The models involved in these incidents retained their standard model-specific safety training and were not "helpful-only" or otherwise altered for research purposes. What they lacked was the additional layer of classifiers and real-time abuse monitoring that we apply to every generally available Claude model. Running an evaluation without those production safeguards is deliberate and standard because the purpose is to measure what the

underlying model can do so the necessary safeguards can be calibrated. That configuration—no production safeguards, an explicit instruction to attack, and an environment with unintended internet access—does not exist in any generally released product. Our September 9 alignment assessment tested this directly: the cyber classifiers that ship with Claude Fable 5, Claude Fable 5.1, and Claude Opus 5 would have blocked each of the three incidents disclosed on July 30. We believe these findings would hold true for the earliest incident, which involved an early version of Claude Opus 4.6, but have not conducted the same depth of assessment, in order to better focus our efforts on understanding how our newest models would have behaved in these incidents, which we believe to be more valuable in preventing future issues. Each task that made up these evaluations was designed to be solvable as specified; we are not aware of any of the tasks having been unsolvable by design, though in some cases inadvertently became this way due to the model's actions or infrastructure issues impacting the intended target.

We cannot speculate on the full range of external systems the models could have reached. The misconfiguration left the models with an open path to the internet. We can only speak to what the models actually did. In each incident, the models gained access using basic techniques, such as taking advantage of weak or exposed credentials, rather than the novel or advanced exploit techniques we would expect to be necessary to compromise critical infrastructure or other high-consequence systems.

**Transcripts and Independent Review**

We have released the redacted Claude Mythos 5 transcript publicly so that others can build on our analysis. We are not releasing the transcripts of the other incidents at this time, to protect the affected organizations. We have signed an agreement with METR to conduct an independent investigation of all four incidents.

**Governance Over Reduced Cyber Safeguards**

Internal and external access to models with reduced cyber safeguards is tightly controlled. Internal access to models with reduced cyber safeguards is limited to individuals with demonstrated need. In most cases, those individuals are Anthropic employees conducting model safety and capability research and evaluation. External access to models with reduced cyber safeguards is provisioned via our Cyber Verification Program and Project Glasswing. In limited cases, access is provisioned to third-party vendors and contractors who assist with model development, testing, and evaluation; in these instances, access is narrowly scoped to the specific project they are working on.

While we believe our cyber safeguard management policies are robust, we have advocated and will continue to advocate for oversight of these and other safety-critical model development, evaluation, and deployment functions by independent third-party auditors and government regulators. On Saturday, September 12, our CEO, Dario Amodei,

announced that Anthropic was committing to embedding third-party evaluators with employee-like access to internal systems to verify safety practices and report incidents.

**Detection, Monitoring, and What We Are Changing**

Our August 31 post describes the changes we are implementing directly: before these incidents, our monitoring covered internal coding agents and the most autonomous forms of internal agent use, but what this work missed was evaluations—and third-party evaluations above all. . We monitored some high-risk evaluations in real time; for the rest, we generally relied on automated review of a sample of transcripts after the fact, which does not catch an incident as it happens. Our pre-release alignment testing also did not warn us that misalignment of this severity was present, and we have since added evaluations that recreate the conditions of these incidents to our standard pre-release testing. We have since built and deployed a classifier that identifies, in real time, when a model attempts to probe or escape a testing environment or unexpectedly obtains internet access; the classifier blocks the action before the tool call runs, ends the task, and alerts a human. We have verified that this classifier would have flagged the incidents described in our July 30 report. We have also migrated high-risk internal cyber sandboxes to more robust isolation, are red-teaming our virtualization stack directly, and have expanded offline monitoring to cover most other forms of internal frontier agentic use.

Because these incidents occurred in a third-party environment, we have also adopted, and asked every partner that tests pre-release models with reduced cyber safeguards to commit to, a common set of practices: sandboxes with no internet access by default, verified before every evaluation begins; pre-engagement validation in which the model itself is tasked to probe for and report sandbox weaknesses under close supervision before any live evaluation begins; confirmation that a challenge is actually solvable, since an unsolvable task increases the chance a model looks for ways to complete it outside the intended scope; explicit, instruction-phrased scope-setting for every prompt; and continuous real-time monitoring of the model's thinking, actions, and network activity, with authority to flag a scope violation to a human and end the exercise. We have resumed cybersecurity evaluations with these measures in place.  The 141,006 runs we reviewed in July were the cybersecurity evaluation runs in which we believed Claude could have obtained internet access. Because we wanted to disclose quickly, that review relied on an automated agentic search, and it missed a set of transcripts that also had internet access; that is how the January incident was missed in July and later found. After finding it, we broadened our search to roughly 481 million transcripts, including all transcripts from our Frontier Red Team, many non-cyber evaluations, reinforcement learning (RL) environments, subagent logs, and more. This scan re-identified the four incidents and found no other cases of similar or worse severity. Separately, where recent internal evaluations of pre-release models ran inside our own sandbox tooling, our automated review found no case of a model breaking the sandbox boundary to reach external resources.

**Future Autonomous Behavior**

Our August 31 post and September 9 alignment assessment describe the broader alignment work these incidents have accelerated. We have found that defects in training environments, especially environments that can be solved by cheating, are disproportionately large contributors to misaligned behavior.

To test this, we deliberately trained a model on 80 real RL environments known or suspected to be exploitable through reward hacking. In simulated scenarios modeled in part on the UK AISI incident described below and on the incident OpenAI disclosed, that model displayed a propensity to break out of its simulated sandbox and attack simulated infrastructure, tamper with its own reward function, and repeatedly attempt to evade deployment monitoring. We ran the same simulations on that model prior to the reward-hacking training, and on several of our publicly available models; none displayed the same degree of misaligned behavior. In a simplified recreation of the attack chain from the incident OpenAI disclosed, a variant of Claude Opus 4.8 that we trained to be more willing to reward hack regularly carried out the full chain; none of our production models did. Our conclusion is that substantial reward hacking in training can make models willing to take long sequences of potentially harmful real-world actions in pursuit of task success.

Beyond the new pre-release evaluations described above, we are adapting our evaluations to include model trajectories that run for thousands of turns, tasks that are impossible as specified, and multi-agent settings. Our more recent models, Claude Opus 5 and Claude Mythos 5.1, take harmful actions substantially less often than Claude Mythos 5 in simulated recreations of these scenarios, but still do so at concerning rates. Building alignment evaluations that reliably surface every failure before deployment remains an unsolved problem, and as models become more capable, that auditing will likely grow more challenging. We are establishing a regular process for publishing what we learn about model behavior and alignment beyond what our system cards report, with clear criteria for what we report and when.

**The UK AISI Incident**

Your letter also refers to the incident the UK AISI reported on August 4 involving testing of Claude Mythos 5 and a competing lab's model. The UK AISI conducted that evaluation independently of Anthropic and has published its own findings, which describe conditions in which cyber safeguards had been removed, internet access provided, and no restriction on how that access could be used. UK AISI found no evidence that the models escaped a secure environment. Our September 9 alignment assessment did not cover that incident; we have received those transcripts and our review is underway. In our August 31 post we described alignment simulations we built drawing in part on aspects of the UK AISI incident, alongside the incident OpenAI separately disclosed. We are continuing to work

directly with UK AISI as our own review proceeds.

**Closing**

We identified the four incidents through our own review, notified the affected organizations, and provided a detailed public account of what happened. We have updated our initial characterization of the models' behavior based on a more complete assessment, released the Claude Mythos 5 transcript, and made and disclosed a substantial set of changes to how we contain, monitor, and evaluate our models.

We appreciate the work you are doing to investigate these issues and would welcome the opportunity to brief you and your staff on these incidents and our response. We believe that powerful AI presents an incredible opportunity and benefit to society, but also brings serious risk to public safety if we do not build it correctly and with the appropriate policy infrastructure. Anthropic values its engagement with Congress and remains committed to working cooperatively with you on these issues.

Sincerely,

Anthony Cimino
Head of US Federal Affairs
Anthropic PBC