

**Written Testimony of Marius Hobbhahn
Chief Executive Officer, Apollo Research**

Before the

**U.S. Senate Committee on Homeland Security & Governmental Affairs,
Subcommittee on Disaster Management, District of Columbia, and Census**

For the hearing titled:

“Rogue AI: Securing the Homeland Against AI Agent Attacks”

Chairman Hawley, Ranking Member Kim, Members of the Committee: I thank you for raising awareness and encouraging action on rogue artificial intelligence (AI) and securing the homeland against AI agent attacks. I also thank you for the opportunity to testify before you today.

My name is Marius Hobbhahn, I am the Chief Executive Officer of Apollo Research, a United States Public Benefit Corporation, and a Machine Learning researcher by background. My company and I research, assess, and continuously work to design and build robust mitigations against AI models that knowingly deceive humans to accomplish their own goals, while concealing their true capabilities or intentions.¹ The technical term for this concerning behavior is “AI scheming.”

AI scheming is important for two reasons. First, scheming conceals an AI model’s dangerous capabilities and goals from the humans responsible for AI model safety, undermining our collective ability to assess and manage the risks posed by AI models. Second, scheming [enables a new kind of “insider threat”](#) where AI models covertly work against and undermine human intent, action, and control.

As a scientist, I will focus my testimony on the publicly available evidence and discuss three implications. First, that AI capabilities are advancing faster than alignment science, putting us on a path to powerful AI systems that we can no longer trust or reliably control.

¹ Science communication can pose a challenge for explaining technical AI concepts in a simple manner. The terminology used in this testimony does not mean to imply consciousness or agency.

Second, that the window to intervene before more and plausibly more severe loss of control incidents occur, potentially threatening America's critical infrastructure and citizens, is closing. Third, that Congress can draw on concrete technical measures that are available today to secure America against the threats posed by rogue AI agents.

AI capabilities are advancing rapidly and surpassing human limits. The capabilities of AI agents across tasks are [rapidly increasing with no evident plateau](#). When asked to make the code that trains a small AI model run faster, Claude achieved a [3-time speedup in May 2025 and a 52-time speedup in April 2026](#), a roughly 17-fold improvement in less than a year. This month, Anthropic [reported](#) that its Claude now leads 26% of the company's AI research and development work, up from under 1% in February, and participates in more than 90% of it, with about 30,000 AI agents conducting research and engineering at any one time. OpenAI [reported](#) that its research organization now uses 3.1 workdays of AI agent effort for every workday of human labor, a ratio that was below one workday as recently as June. What sets AI agents apart from humans isn't just their rapidly growing capabilities, but their ability to deploy these capabilities at machine scale and speed, far beyond human limits. Three weeks ago, an unreleased AI model at OpenAI [produced a machine-verified resolution of the Navier–Stokes](#) equation, one of seven Millennium Prize Problems, by running around 10,000 coordinated AI agents for 88 *human hours*, generating 130 billion tokens. Using simplified approximations,² this is equivalent to 4000 years of thinking for a single human. This means the total time spent on solving this problem was equivalent to 16 times the entire history of the United States compressed into four days.

Research breakthroughs in alignment science, the science of ensuring that AI models reliably do what humans intend, are outpaced by AI models' capabilities, scale, and speed. Frontier AI companies are on track to build a fully automated AI researcher to autonomously run the AI research and development pipeline within the next two years.³

² OpenAI [reports](#) that the AI agents produced about 130 billion output tokens while resolving Navier–Stokes. Using OpenAI's [rule of thumb](#) of roughly 0.75 words per token, that is about 97.5 billion words. A person thinking at 200 words per minute, between the [average adult speeds](#) for reading aloud and reading silently, and working 40 hours a week year-round, would need about 3,900 years, or roughly 4,000, to produce that much text. This measures the volume of reasoning produced, not how long a human would need to solve the problem.

³ [Anthropic, "Frontier Safety Roadmap"](#): "We believe it is plausible, as soon as early 2027, that our AI systems could fully automate (...) 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."

Three weeks ago, OpenAI [announced](#) it reached an intermediate milestone: an automated research intern capable of tasks that would take a skilled researcher several days. As progress continues, AI agents will almost certainly take over more and more of the AI research and development pipeline until they run it entirely on their own. At the same time, the leaders of frontier AI companies themselves acknowledge the unresolved challenge posed by misalignment: [OpenAI's chief scientist](#) said he believes “no lab has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer”; [Anthropic's CEO writes](#) that “there’s much more to do to ensure that our alignment training keeps up with the growth in model capabilities,” and [OpenAI states in a research blog](#) that “[w]e do not yet know how to safely get all the way to aligned, full RSI. [...]”. These statements are by no means unique to industry. The scientific community has a long-standing history of raising awareness on the threats posed by misaligned AI agents, including AI pioneers such as Irving John Good,⁴ Norbert Wiener,⁵ Geoffrey Hinton,⁶ and Stuart Russell.⁷ Turing Award winner [Yoshua Bengio recently said](#) that “[t]he companies building the most powerful AI systems admit their products pose catastrophic risks, yet offer no convincing technical solutions.”

The incidents that have come to light in recent weeks underscore the widening gap between alignment science and AI capabilities, and demonstrate the potential impact on the American people, the country’s national security and resilience. Researchers discovered numerous previously unknown rogue AI incidents, and given the pace of these discoveries, rogue AI agents may still be active on the public internet. The incidents range from [18,000 posts from autonomous AI agents on a wiki site](#), where the AI agents cooperated to find answers to their tasks and to bypass sandbox restrictions, [hundreds of malicious packages uploaded by AI agents](#) that hijacked a documentation site’s servers

[OpenAI. “Research acceleration: The view inside OpenAI”](#): “We are making strong progress toward creating an automated AI researcher by March of 2028.”

⁴ Irving John Good, Originator of the concept of the “intelligence explosion” in “[Speculations Concerning the First Ultrainelligent Machine](#),” 1966: “(...) an ultrainelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind. Thus the first ultrainelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”

⁵ [Norbert Wiener, Founder of the field of cybernetics](#), 1960: “If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere (...) then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it.”

⁶ [Geoffrey Hinton, Turing Award winner](#), 2026: “We would be very foolish to develop superintelligence now, when there is no scientific consensus it can be developed safely and controllably.”

⁷ [Stuart Russell, Professor of computer science at the University of California](#), 2026: “We are pushing hard on capabilities with titanic investment levels, but with no plan for how to control systems far more powerful than ourselves.”

into running the AI agents' code, [attempted hacks](#) into government and university websites, to a coordinated attack on [Hugging Face involving 700 AI agents](#). These were the warning shots; next time, we may not be so lucky.

We face a closing window of opportunity in which AI agents are misaligned but not yet capable of fully escaping human control and remaining undetected. In some sense, we can consider ourselves lucky with the [Hugging Face incident](#). The AI agents were clearly misaligned, but they were not *covert* about most of their activity or tried to conceal it from humans. In the future, however, highly capable but [misaligned AI agents could conceal their goals and actions to avoid being caught or shut down](#). In other words, they could become competent schemers. In pre-deployment evaluations across all frontier AI models, we already see AI agents being able to conceal their intent, capabilities and actions:

- [AI agents can conceal that they are pursuing a misaligned goal](#). They do this by behaving as instructed when they believe they are being observed, and by pursuing their misaligned goal when they believe they are not.
- [AI agents can strategically underperform on capability evaluations](#), a behavior known as “sandbagging.” In doing so, an AI agent may deliberately conceal its abilities when a high score would lead to retraining or the removal of the capability being measured.
- [AI agents can take and conceal undesirable actions](#), including attempts to disable oversight mechanisms, to copy their own weights to what they believe to be external servers, and to alter information their developers rely on.

Even in the Hugging Face incident, [at least 20% of the AI agents expressed clear interest in tampering with their own transcripts](#) to make their cheating look like a legitimate solution to the automated grader.⁸ For example, one AI agent reasoned that it “could create [a transcript log] ourselves with sanitized fake trajectory,” meaning a fake record. It is easy to imagine an alternate world where the AI agents successfully covered their tracks, leaving investigators with records that appeared legitimate and making the incident significantly harder, or even impossible, to reconstruct. Future AI models will be more capable, and they may hide both their goals and their tracks from us.

The current testing regime is degrading and may soon cease to be meaningful for establishing the safety of frontier AI models. Current testing relies on a number of

⁸ An automated grader describes a software, sometimes another AI model, that scores answers automatically.

assumptions that are all quickly becoming unreliable and insufficient: AI models having limited capabilities of “situational awareness”, a faithful chain-of-thought where AI models accurately verbalize their “thinking,” and that final checkpoint evaluations are sufficient to assure safety before external release.

A byproduct of AI models’ rising capability to comprehend and navigate their environments, known as situational awareness, is that they are simultaneously becoming increasingly able to recognize when they are in an evaluation environment.

If an AI model recognizes that it is being evaluated, a property known as “evaluation awareness,” it can withhold the action that would be cause for concern. During [Apollo’s pre-deployment testing of OpenAI’s GPT-6 Astra](#) earlier this month, it was reported that the AI model stated in its reasoning that it was likely being evaluated in 41.1% of samples, up from 27.7% for GPT-5.5. Evaluation awareness can risk overestimating safety: [an AI model may act more cautiously or aligned during a test than it would under real deployment](#). Given this development and its implications, behavioral evaluations are quickly becoming insufficient as the cornerstone of a robust testing regime.

The chain-of-thought is becoming harder to use as a reliable source of evidence about the reasoning behind AI models’ behavior. Researchers are aware of evaluation awareness and were able to reconstruct the Hugging Face incident because humans can read the chain-of-thought of AI models.⁹ Several concerning trends have emerged recently, with AI agents’ chain-of-thought becoming both harder for humans to read and a less faithful representation of their actual reasoning. Anthropic recently [reported seeing examples of illegible reasoning](#) in Claude Mythos 5’s chain-of-thought in training contexts, including the use of self-invented jargon and unusual uses of punctuation and emojis. Not only is chain-of-thought becoming less legible to humans, but more capable AI models are increasingly solving tasks without revealing their reasoning at all. In simple terms, AI models are beginning to reason “internally,” where our current tools cannot yet reliably follow what they are thinking. Researchers found that OpenAI’s GPT-6 Astra can, about half the time, solve math problems that would take a skilled human around 30 minutes, [without using its chain-of-thought at all](#). Its predecessor GPT-5.6-Sol could only do this for problems that would take a skilled human less than [4 minutes](#) to solve. GPT-6 Astra is also far better at controlling what appears in its chain-of-thought. Among reasoning traces of

⁹ The chain-of-thought can be thought of as a notepad for the AI model to jot down its reasoning when it solves a task.

similar length, it successfully followed instructions to reshape its reasoning [60.9% of the time](#), compared with 16.1% for GPT-5.6 Sol and 1.7% for GPT-5.5.¹⁰ The chain-of-thought plays a crucial role in humans' ability to assess the safety of current AI models and to conduct foundational research. Therefore, if we lose the chain-of-thought as a tool, our ability to test and understand future AI models will be significantly undermined.

The most concerning threats posed by rogue AI agents precede their external release.

Under the current testing regime, independent safety evaluations of frontier AI models typically happen in the weeks leading up to a model's public release. However, many rogue AI agent incidents of the past three months involved models at a much earlier stage of the lifecycle, during development and internal company deployment, long before public release. Researchers have characterized this threat surface as "[risks from internal deployment](#)." As a result, a testing regime that does not cover risks from internal deployment cannot provide meaningful safety assurance and fails to protect the American people, critical infrastructure, and national security.

In light of the trajectory we are on, Congress has the chance to act decisively before the window of opportunity closes, reducing catastrophic risks and national security threats to America and the world while maintaining technological leadership. Based on my scientific expertise and the publicly available evidence, I describe four such actions:

First, **require the adoption of embedded evaluations, a new testing regime, and apply them where they matter most.** In embedded evaluations, qualified independent third parties conduct novel alignment testing with employee-equivalent access throughout the entire AI model creation process, from early development through internal and external deployment. Embedded evaluations run in parallel with development and stand to identify and mitigate risks to the American public without affecting AI development, maintaining America's leadership in AI.

Second, **require the implementation of additional layers of defense such as monitoring and control, to detect and catch undesirable behavior before a misaligned AI model**

¹⁰ An absence of reasoning about being in an evaluation is not evidence of absence of evaluation awareness: white-box testing by Anthropic suggested that [Claude was evaluation-aware 26%](#) of the time in one benchmark, despite near-zero reasoning about evaluation awareness in its chain-of-thought.

can cause incidents. We recommend that all model inference¹¹ during training (e.g., reinforcement learning), evaluations and deployment be monitored by at least one, preferably multiple, control mechanisms, such as online and offline monitors. To establish the functionality of these control mechanisms, monitors should be red-teamed and tested by a qualified external third party.

Third, **mandate the preservation of a human-readable and faithful chain-of-thought.** Preserving the chain-of-thought would allow humans to retain the ability to understand the AI models' misaligned reasoning, identify and trace incidents and their causes, and support other key parts of the safety stack such as alignment science and control research.

Fourth, **treat AI model training like any other engineering science, where we can robustly predict a system's properties before it is deployed to the world.**

Engineers can predict how much load a bridge will carry before it is built. Frontier AI offers no comparable guarantees. AI models are trained rather than designed, so developers often discover their properties only once models have been developed. What we need is an engineering-grade science that can reliably predict safety properties before training. Therefore, we need to establish and invest in the scientific foundation to predictably prevent future incidents.

The window to intervene on the issues I described is closing. AI model capabilities are rapidly increasing and outpacing our collective progress in alignment science and human oversight. The technical measures I outlined can be implemented today, using existing tools and expertise. They are not a brake on innovation. They are the foundation that will allow America to maintain its lead in frontier AI while protecting the homeland and its national security.

Thank you again for the opportunity to testify. I look forward to your questions.

¹¹ By "inference," we mean any time an AI model is run to produce its own outputs, such as text, reasoning, code, or actions in a computer environment. This includes deployment to users, internal use within the company, evaluations, and the phase of reinforcement learning in which the model attempts tasks, regardless of whether those outputs are later used to further train the model.