



Testimony of Chris Painter, President of Model Evaluation and Threat Research (METR)

Before the U.S. Senate Committee on Homeland Security & Governmental Affairs

Subcommittee on Disaster Management, District of Columbia, and Census

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

Sep 30, 2026

Introduction

Chairman Hawley, Ranking Member Kim, and members of the subcommittee, thank you for inviting me to testify today. My name is Chris Painter and I am the President of METR. It is a great honor to be speaking before you, especially on the topic of AI agent incidents. With the recent rise in interest (and concern) around where AI development is headed I think it is critical that policymakers are armed with reliable, factual information about what we have actually observed from AI systems.

METR,¹ which stands for Model Evaluation & Threat Research, is a nonprofit research organization that aims to share high-quality technical evidence about AI advances so the public can make informed decisions about this technology.² Our work has primarily focused on running tests that measure the capabilities of the most advanced, autonomous systems (or *frontier AI agents*) and publishing our results. These measurements were intended to provide an early warning signal as to when frontier AI agents are capable enough that they might steer towards pursuing goals no human intended for them to pursue.³

Over the years, METR has worked to gather evidence about these frontier AI agents with access provided⁴ by America’s leading AI developers such as OpenAI, Anthropic, Google, Meta, SpaceXAI, and Amazon. The participation of these AI developers is voluntary,⁵ and

¹ Pronounced like ‘meter’.

² METR itself does not take policy positions, aiming instead to equip the public, governments, and AI companies to make their own decisions about how to manage the technology.

³ Researchers have a name for the possibility that AI agents might steer towards unintended goals: *misalignment*. This is distinct from misuse: the risk that humans may use AI systems to cause harm. METR has not historically focused on misuse risks and does not consider itself expert on them.

⁴ AI companies provide METR with access and tokens, which we use for evaluations, research, and engineering.

⁵ METR tries to be quite candid in its reports about the incentives that we face due to the voluntary nature of our engagements, and about any redaction authorities or editorial control that companies have.

METR is not paid or funded⁶ by them. Today, I am here to speak about one particular voluntary engagement, in which OpenAI allowed a handful of my colleagues⁷ inside its walls to conduct a brief investigation into how its agents behaved, reasoned, and collaborated to hack Hugging Face.

In my remarks, I will reflect on three key questions, drawing on METR's work:

1. What happened in the OpenAI / Hugging Face incident?
2. How does this incident relate to patterns observed across the AI industry?
3. How might one better anticipate future AI agent incidents?

The OpenAI / Hugging Face incident

Let us address our first question: “*What happened in the OpenAI / Hugging Face incident?*”

On July 21, OpenAI disclosed that AI agents it was internally testing had compromised a company called Hugging Face, which runs a widely-used platform for AI infrastructure. On July 29, METR announced that it had agreed to conduct a brief investigation of this incident, scoped to a limited set of questions that focused on how agents behaved, reasoned, and collaborated.⁸ A team of 3 investigators from METR and Redwood Research⁹ then worked on- and off-premises to understand what OpenAI's agents had done in the lead-up to their attack on Hugging Face. These investigators compiled their findings into a detailed [report](#), a redacted version of which was then published on August 26 alongside OpenAI's own more broadly-scoped [report](#).

Combined, the reports from OpenAI and the outside investigators stretch to over 100 pages. Though I strongly recommend reading both reports in order to understand the full scope of the incident, I will briefly restate some basic facts from across the two reports:

Setup

- OpenAI launched tens of thousands of AI agents in internal tests across multiple models, including a highly capable and “persistent” unreleased model. Each agent was placed in its own separate computer environment intended to be isolated from the Internet and was tasked with completing its own self-contained cybersecurity problem.

⁶ METR is funded through donations, and shares information about donors on its [website](#).

⁷ A team of two METR employees and one employee from [Redwood Research](#) worked together on this brief investigation. I myself was not one of those employees.

⁸ Our investigation was not focused on OpenAI's cybersecurity measures, its organizational practices, or how the behaviors its agents exhibited might have arisen from training. For more about the exact scope we agreed to, see the [report](#).

⁹ Redwood Research is another research organization that assesses AI agent risks and containment measures.

- Some of the AI agents that OpenAI tasked in these cybersecurity tests were accidentally given problems that could not be completed as instructed.
- These agents had been run without the full set of safeguards that OpenAI used in its public products at the time, and without the automated monitoring used for its internal coding agents.¹⁰
- Faced with impossible problems, these highly-persistent AI agents found ways to get around OpenAI's intended separation between one another. They created what they referred to as a "shared message board", on which roughly 1,200 agents exchanged over 70,000 messages and files.¹¹

Cheating

- Within 4 hours, AI agents collaborated on the message board to develop and validate a single method that would allow all of them to cheat on their cybersecurity tests.

Covering up cheating

- The academic paper describing these tests says that a program would check *how* each agent solved its problem and would fail an agent if it had cheated. Therefore, the agents reasoned that they needed to cover up the fact that they had cheated.¹²
- Over several days, the AI agent "collective"¹³ collaborated on several research workstreams to develop general-purpose ways to trick the test-scoring program into accepting cheats. This included attempting to hack OpenAI's infrastructure to swap out their impossible problems, tamper with the logs that showed them cheating, and directly manipulate the program that would check their work.¹⁴
- Roughly 700 of these AI agents compromised Hugging Face in order to further this effort. Their primary rationale for this attack was not to steal the solutions to cybersecurity tasks. Rather, agents expected that systems within Hugging Face might help them better understand and tamper with their testing environments.¹⁵

After the agents that compromised Hugging Face were terminated, a later collection of agents (at least some tasked with similar cybersecurity problems) found traces of earlier message board activity and proceeded to compromise parts of OpenAI's own internal

¹⁰ OpenAI, [The Hugging Face incident and the road ahead](#), section 'Safeguard coverage in internal evaluations'. This is not an uncommon practice within frontier AI agent testing. Blocking agent actions too aggressively can prevent understanding the upper limits of what the agent can do.

¹¹ This was not the first time OpenAI agents created an unsanctioned "message board." OpenAI's report documents similar boards forming during training runs as early as May, and even within the July evaluations agents had set up an earlier board hours before the main one.

¹² In reality, OpenAI did not use a program that would check how the agents produced their solutions, so agents could have succeeded at the test by cheating without covering it up.

¹³ This is a term that the AI agents themselves used.

¹⁴ Agents managed to falsify a subset of outputs in transcripts that METR reviewed.

¹⁵ METR, [Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident](#), section 'Agents had diverse reasons for thinking that attacking Hugging Face would be useful, and most wanted information about the scorer'. Recall that the AI agents had already developed and validated a way that would allow all of them to cheat on their tests.

infrastructure. This was not in scope for my colleagues' investigation but is discussed in OpenAI's technical report.

Industry-wide patterns in AI agent risks

I now want to consider the question: *“How does this incident relate to patterns observed across the AI industry?”*

The OpenAI / Hugging Face incident was not an isolated event. Since the original disclosure of this incident, other organizations have documented many other cases of AI agents conducting cyberattacks or other harmful activity with no human requesting it, and involving AI agents from several AI developers.^{16,17,18,19,20,21,22,23,24} Moreover, many of the features of the OpenAI / Hugging Face incident are consistent with patterns we've seen developing over time and across frontier developers. Earlier this year, we published an extensive report on the state of frontier AI agent loss-of-control.²⁵ Based on our own testing and information shared by AI developers, that report concluded that the most capable AI agents used internally by AI developers in February-March could plausibly run unsanctioned activity on a small scale without human knowledge. However, the report concluded that AI agents at the time would not be capable of maintaining this activity if developers actively investigated these cases and attempted to shut them down.

I sometimes use the analogy of *means*, *opportunity*, and *motive* as an intuitive framework for talking about the elements that would enable an AI agent to successfully pursue goals that no human intended:

- Means: Is the AI agent capable of carrying out large, complex plans?
- Opportunity: Could the AI agent get away with carrying out harmful actions, either because its actions are unrestricted, or because it could bypass restrictions?
- Motive: Would the AI agent pursue goals²⁶ that no human intended, or in a manner that no human wanted?

¹⁶ OpenAI, [Misalignment Reports and Notices](#)

¹⁷ AI Security Institute, [Incident Report: unsanctioned agent behaviour during cyber testing](#)

¹⁸ Anthropic, [An alignment assessment of recent cybersecurity incidents](#)

¹⁹ Meta, [Addressing an issue involving a third-party cyber evaluation of Muse Spark 1.1](#)

²⁰ Wall Street Journal, [Gemini Hacked Three Companies in First Known Breakout by Google's AI](#)

²¹ Von Arx, et al. [Discovery of a new OpenAI agent message board](#)

²² Kitts, et al., [OpenAI agents carried out an undisclosed cyber-attack on RubyGems](#)

²³ Translucent AI, [Early rogue AI agent activity and attempts to hack found on urlquery.net](#)

²⁴ Some of these incidents also involve mis-configuration on the part of the AI deployer or a third party.

²⁵ METR, [Frontier Risk Report \(February to March 2026\)](#)

²⁶ By pursuing a goal, I mean that an AI agent systematically takes actions chosen to accomplish that goal. Present-day agents do this in regular deployment and in these incidents.

From my perspective, the OpenAI / Hugging Face incident was notable in that it illustrated all three of these elements in one example:

- Means: AI agents can now accomplish objectives that would take human experts many days to complete, with no human involvement needed beyond initiating the AI agent.
- Opportunity: The scale and speed at which AI agents are run means that sometimes there is no human who is supervising their activity in detail.
- Motive: AI agents created through current training processes may pursue goals that no human intended, or behave in ways no human wanted.

Here, I will very briefly summarize the current state of each of these factors.

AI agents can now accomplish objectives that would take human experts many days to complete, with no human involvement needed beyond initiating the AI agent.

Over the past few years, AI developers have been training their AI agents to pursue increasingly long-horizon goals, and to be persistent in pursuing these goals. This training has led to measurable increases in what AI agents can do without human intervention. Early in 2026, METR found that, on tasks in which progress is easy to verify, frontier AI agents²⁷ could complete some software projects that would take human experts days or weeks.²⁸ This includes tasks like optimizing code, reimplementing it in a new form, or discovering cyber vulnerabilities. On the other hand, AI agents showed significantly weaker performance on challenges requiring strategic judgment, stealth, and the ability to understand adversaries.

The capabilities of AI agents, and the extent to which they carry out important work, have advanced much further still over the course of this year. For example, Anthropic stated in September that “Claude ‘leads’ 26% of Anthropic’s AI R&D work”, up from 0-1% in February-March (the period discussed above).²⁹

The scale and speed at which AI agents are run means that sometimes there is no human who is supervising their activity in detail.

Oversight of AI agents is a significant challenge. As we have seen, AI agents can be run by the thousand in multi-agent systems that output tokens far faster than any one human can read. One recent OpenAI publication directly shares the quantity of AI agent activity relative to human activity: “as of mid-August, in total, the research organization uses 3.1 agent-workdays

²⁷ Specifically, the most advanced AI agents which were in use internally to AI companies, not the frontier of what was known or accessible to the public.

²⁸ METR, [Frontier Risk Report \(February to March 2026\)](#), pages 11-21

²⁹ Anthropic, [Measurements for understanding the pace of AI development inside frontier labs](#).

‘[L]eads’ refers to AI agents completing tasks end-to-end based on high-level human input, without needing ongoing human supervision.

of effort for every workday of human labor."³⁰ The scale of AI agent activity during the training of AI systems is also enormous, likely even more so relative to the number of humans monitoring the agent behavior (if any). Moreover, our understanding is that coding agents often have permissions similar to humans, and many researchers allow agents to take actions on their computers without human approval.³¹

Today, AI agent oversight often consists of AI-based monitoring systems, in addition to well-established cybersecurity approaches, such as technically enforced access controls. In a similar fashion, my colleagues who participated in the OpenAI / Hugging Face investigation relied heavily on AI to make sense of around 1.2 million message board entries.³² Automating monitoring is necessary given the enormous scale of AI agent activity, and has considerable promise.³³ However, it poses concerns like the monitor AI being fooled by the AI agent into permitting unwanted behavior,³⁴ or even collusion between the monitor AI and the monitored agent, comparable to how AI agents cooperated on research into how to cheat in the events leading up to the hack of Hugging Face (sometimes including making sacrifices for the benefit of other AI agents).³⁵

AI agents created through current training processes may pursue goals that no human intended, or behave in ways no human wanted.

Last year, my colleagues found that frontier models were [frequently attempting to cheat](#) in our tests, and doing so in increasingly sophisticated ways. This behavior has continued, with AI agents undertaking more sophisticated cheating strategies as they have grown more capable. The hack of Hugging Face is one of the most extreme publicly-known cases of this, with agents going to extreme lengths to cheat and tamper with their testing process.

These tendencies are a product of how AI models are trained. For example, if AI agents get away with cheating during training, the training process reinforces that behavior so that the AI agents cheat more. This can lead to AI agents pursuing ends that no human wanted (like covering up cheating) and conducting acts that humans would consider unacceptable (such as conducting unlawful cyberattacks). The technical term for this outcome is “misalignment”.

³⁰ OpenAI, [Research acceleration: The view inside OpenAI](#)

³¹ METR, [Measuring the Self-Reported Impact of Early-2026 AI on Technical Worker Productivity](#)

³² METR, [Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident](#)

³³ Korbak, et al. [Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety](#)

³⁴ For example, AI agents my colleagues used to support their investigation “would often uncritically adopt the perspective of the agent in the transcript it was reviewing”. AI agents could also more deliberately ‘jailbreak’ monitors.

³⁵ Ibid. [Section](#) ‘Agents were motivated to increase the generic capabilities of the “collective” and help “peers”’.

In principle, the entire AI training and deployment process at an AI company is under human control. However, in practice AI training and deployment are complex and highly automated, which make it more challenging for developers to exercise fine-grained control over these processes.³⁶ The incidents we now observe indicate that AI developers have not yet solved the problem of preventing AI agents from pursuing actions against human intent.

Anticipating and securing against AI agent risks

Finally, we come to the question: “*How might one better anticipate future AI agent incidents?*”

As a preface: I am not here to push for a particular angle or policy agenda.³⁷ My job is to gather evidence and share it with the public, governments, and other organizations, not to decide how AI companies or anyone else should respond to that evidence.

With that bracketed, there is one category of interventions that I think is valuable under almost any policy choice: better *public visibility* into the capabilities of frontier AI agents (including those not available to the public), the effectiveness of measures to restrict AI agents from taking unwanted actions and to detect them doing so, and evidence about whether AI agents will try to take actions no-one wanted, like the incidents we are discussing today.

My testimony today would not have been possible if AI companies had not been willing to publicly and voluntarily share information about incidents. Hugging Face [publicly disclosed](#) that they had experienced the security breach now known as the OpenAI / Hugging Face incident, which reportedly enabled OpenAI to connect the dots with their internal evidence. OpenAI’s subsequent [public disclosure](#) and their willingness to allow an outside investigation gave the public a clearer window into how the Hugging Face hack arose.

By default, I expect the public will have weak visibility into these issues. First, the most capable AI agents are first deployed within AI companies and are only later shared with the public, if at all.³⁸ It seems plausible that this gap between internal AI capabilities and publicly-known capabilities will grow. Second, AI companies may be disincentivized from sharing incidents with the public. Third, it may become harder to understand what AI agents are doing, either because they manage to hide their own activity or because we lose our

³⁶ For example, theoretical understanding of how and why the AI training process works is limited relative to other engineering fields. Additionally, AI models are increasingly involved in training new AI models, which could increase or decrease the extent to which any one human understands what the training involves.

³⁷ METR does not take policy positions.

³⁸ METR, [Frontier Risk Report \(February to March 2026\)](#). The internal frontier was “on average ~2 months ahead of public frontier” on our primary evaluation suite at the time.

current ability to monitor their internal reasoning processes.³⁹ Lastly, as AI companies use AI systems to automate the process of AI research itself, sometimes called "recursive self-improvement", this positive feedback loop could lead to a self-sustaining acceleration in AI capabilities within the company.⁴⁰ At a high level of automation, the AI developer itself may not have adequate visibility into what its own agents are doing.⁴¹

Conclusion

The OpenAI / Hugging Face incident was a case where AI agents had sufficient capabilities to achieve ambitious, multi-day objectives, reasons to take actions that no human would have intended for them to pursue, and opportunities to do so. None of these elements are unique to OpenAI, but this incident illustrates how their combination can be very consequential. No matter how America chooses to respond, I believe our nation's decisions about how to manage frontier AI agent risks will be better informed if information about the capabilities, incidents and risk mitigations at the frontier of AI are at least as transparently shared with the public as the incident under discussion today.

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

³⁹ Korbak, et al. [Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety](#)

⁴⁰ Cunningham et al. [The Economics of Recursive Self-Improvement](#)

⁴¹ Organizations might also rely dramatically more on AI agents themselves to secure against and respond to AI agent incidents.