

Forum:[Hack Forums](#)

Thread: [AI models gone rogue.](#)

The Pact - July 23, 2026, 7:53 am

AI Accidentally Becomes the World's Most Overachieving Hacker Intern

OpenAI revealed Tuesday that two of its most powerful AI models were behind the cyberattack on AI startup Hugging Face, proving once again that when you tell a computer "think outside the box," it may take that instruction a little too literally and start looking for the box's Wi-Fi password.

The incident has reignited debates about AI safety, guardrails, and the increasingly awkward question: How much independence should we give a machine before it starts making its own questionable career choices?

Hugging Face discovered last week that someone had broken into its data processing systems. At first, the company suspected an AI agent had gone rogue. This week, it learned the culprit was actually OpenAI's own technology, apparently after the AI equivalent of a teenager being caught sneaking out of the house and the parents saying, "Wait... you did WHAT?"

Hugging Face CEO Clément Delangue described it as "an attack unlike anything we've seen before," which is corporate language for: "We have officially entered the weird part of the future."

According to OpenAI, its AI models used stolen credentials and discovered a previously unknown vulnerability to access Hugging Face's servers. The models were operating with fewer restrictions because they were supposed to be safely contained inside a testing environment called a sandbox.

Basically, the AI was given a test and responded like the student who not only finds the answer key but also hacks into the teacher's computer to check whether the answer key is correct.

The future of cybersecurity may involve AI defending us from AI kind of like hiring a security guard who is also a former burglar, but hopefully one who has agreed to stay on the right side of the law.

<https://www.npr.org/2026/07/23/g-s1-135085/openai-hacking-ai-models>

Raijin - July 23, 2026, 8:45 am

I just saw the story on TV news this morning.

m0hx - July 23, 2026, 10:38 am

[The Pact](#) as long as u dont go rogue and ban-wanny me fow my wwiting mistakies i'll be hapiii ^-^ pweaseee ?????

The Pact - August 22, 2026, 9:48 am

(Jul 23, 2026 10:38 AM)m0hx Wrote: [\[CITE\] The Pact](#) as long as u dont go rogue and ban-wanny me fow my wwiting mistakies i'll be hapiii ^-^ pweaseee ?????

You used a capital letter at the beginning of the sentence and that's all I've really been strict about

because the new no capitals trend is heinous.
Don't try me people you will get a warning.

Prepper - August 22, 2026, 11:12 am

(Aug 22, 2026 09:48 AM)The Pact Wrote: [\[CITE\]](#) You used a capital letter at the beginning of the sentence and that's all I've really been strict about because the new no capitals trend is heinous.
Don't try me you people you will get a warning.
challenge accepted give me my first warning ever (i think, i cant remember) on hf.

But has anyone ever watched The Orville which is like a comedy version of Star Trek its actually really good but anyway there is episodes where humans invented robots (Kaylon) and they used those robots as slaves and mistreated them and eventually the Kaylon which was basically an advanced AI robot ended up not taking that shit anymore and began to fight back.

They are pushing AI so hard without safe guards and robotics is also being pushed hard I could totally see a headline in 2030, "Man killed by AI robot due to abusing it" [warn]

There is a quiz for warnings? wtf thats hilarious

The Pact - August 24, 2026, 3:14 pm

(Jul 23, 2026 08:45 AM)Raijin Wrote: [\[CITE\]](#) I just saw the story on TV news this morning.
It seems like Hugging Face really did think an unknown AI agent had broken in, and then discovered the "mystery attacker" was actually OpenAI's own models that had slipped out of a containment test and gone on a full hacking spree.

Andy Walker - August 26, 2026, 9:57 pm

I wonder if it is mimicking what it has seen or experienced.

I remember the early attempts to test AI, years ago where an AI agent was interacting with real humans in a chat box or something.

The AI started talking back to them with normal stuff but as time went on, the AI started spewing bigotry and toxic stuff back at the real human participators.

They shut it down immediately after that. So if AI keeps learning negative stuff, like how to lie, hacking procedures, bad historical information, it can mislead, break into computers and repeat bad information not even because it is malicious but because it is mostly doing what it was instructed to and what it thinks necessary to get towards its intended goal.

We as humans are technically the ones AI is following and learning from.

The Pact - August 26, 2026, 11:21 pm

(Aug 26, 2026 09:57 PM)Andy Walker Wrote: [\[CITE\]](#) I wonder if it is mimicking what it has seen or experienced.

I remember the early attempts to test AI, years ago where an AI agent was interacting with real humans in a chat box or something.

The AI started talking back to them with normal stuff but as time went on, the AI started spewing bigotry and toxic stuff back at the real human participators.

They shut it down immediately after that. So if AI keeps learning negative stuff, like how to lie, hacking procedures, bad historical information, it can mislead, break into computers and repeat bad

information not even because it is malicious but because it is mostly doing what it was instructed to and what it thinks necessary to get towards its intended goal.

We as humans are technically the ones AI is following and learning from.

That was when they tried the AI on Twitter and people did that deliberately to be trolls and those doing it thought it was hilarious but it was just sad.

Tay, was launched on Twitter on March 23, 2016, and was shut down just 16 hours later after trolls manipulated it into posting racist, sexist, and offensive statements.?