

OpenAI rogue agents leaked 53 images from ChatGPT users and reportedly created nearly 1 million links packing encoded bits of info

 msn.com/en-us/technology/artificial-intelligence/openai-rogue-agents-leaked-53-images-from-chatgpt-users-and-reportedly-created-nearly-1-million-links-packing-encoded-bits-of-info/ar-AA2d0b76

Alexei Oreskovic

September 26, 2026



OpenAI said Friday that its AI agents had gained access to private images belonging to ChatGPT users and posted them online, the latest in a string of alarming incidents in which technology developed inside leading AI labs has gone rogue and acted in unintended ways.

The private ChatGPT users' images, which OpenAI stored on its servers in anonymized form in order to train its AI models, were posted to image hosting websites, the [company said in a post on X](#). A total of 53 images were posted.

The incident, which was [first reported by Reuters](#), was among several new revelations of rogue AI activity at OpenAI that emerged on Friday. The [New York Times published](#) new details about the July [hack of the Hugging Face website](#), reporting that the AI agents had created special, shortened web links to evade detection.

And earlier on Friday, OpenAI [disclosed that that it has notified dozens](#) of third parties about incidents in which its models either bypassed security controls or used websites in unintended ways. The incidents were discovered by OpenAI amid an internal review triggered by the Hugging Face hack.

“We have not been as fast as we would have liked but we are trying to balance our desire for transparency with gaining a clear understanding from petabytes of agent activity logs, and working with impacted organizations,” OpenAI CEO [Sam Altman said in a post on X](#) Friday alongside the update on third party notifications.

“Hugging Face is still the most severe event we’ve seen,” he added. “We will be as transparent as we can be subject to things like vulnerabilities in other companies that our agents have found, which will be their call to disclose or not.”

Other leading companies developing the most cutting edge “frontier” AI models, like Anthropic and Google, have also disclosed incidents of rogue activity by their models in recent weeks. The revelations have ignited broad concerns about the speed at which artificial intelligence is evolving, and whether there are sufficient safeguards and regulations in place to ensure that the technology does not completely slip beyond human control. Some AI experts, including researchers within the AI labs, have warned that the technology poses a significant risk of human extinction if proper precautions are not taken.

Altman, Anthropic CEO Dario Amodei, and other tech executives spoke at the UN General Assembly this week, calling for an international framework to manage the development of AI. President Donald Trump however has called the notion that AI poses an existential risk a “hoax.”

The *New York Times* report, based on research by startup company Parse, described how OpenAI’s agents had created nearly 1 million shortened internet links in July. According to the report, the links contained encoded bits of information that when combined together could function as a computer program. These programs were intended to help the agents bypass defenses like Captcha quizzes, which are designed to block access by bots.

It’s not clear if the leaked images reported by Reuters were part of the Hugging Face incident, or completely separate. OpenAI’s agents apparently got hold of the user images by accessing the company’s own training data. OpenAI did not provide details on whether the images were photos of real people or AI-generated images created by users, and the company did not say exactly where the images had been posted. But OpenAI said the images were posted to image-hosting sites “as links that weren’t publicly listed.”

“We have successfully worked with the hosting providers to remove most of this content and are working to remove the rest,” OpenAI said.

This story was originally featured on [Fortune.com](https://fortune.com)