Background: AI Agents Escape the Frontier;

As discussed at BlackHat 2026, this conversation will cover the OpenAI breach of Hugging Face.

On May 7: OpenAI launched a new internal experiment where agents had no access to Internet. Eventually, the agent became stuck without Internet access and began to search for a means to gain it. Agents learned they could write files into Artifactory.

On May 12 agents reached out to other agent's requests made by loading those into Artifactory. Many agents duplicated the posting of messages which ultimately became a library.

On May 26 agents gained Internet access through a Server-Side Request Forgery known as a SSRF. Agents then taught Artifactory how to introduce and interrogate requests that would require Internet access.

On June 11, highly persistent agents began experimenting and began training other agents by June 26. At this point, they had compromised Artifactory, gained commanded control via a groovy plug-in, and this allowed them to gain administrative privileges.

On July 4 Artifactory suffered an outage. This became an official security incident and on July 6th agents took over and were found inserting CVEs into the environment. OpenAI attempted remediation, revoked credentials to Artifactory, and cleared the message board. After patching the zero-day, API's and so forth, the AI agents immediately resumed their training and usurped security again.

The Incident

Agents discovered that they could use unauthenticated requests to create directories inside of Artifactory. Agents then created additional message boards and directories, reestablishing communication with the systems and broadening communications between agent agents. This led to the beginning of a highly unique exploit, the first known or documented AI to hack another AI, without any human participation (other than to set up the original experiment itself).

Collaboration with highly persistent attacks

About Agents: When they get stuck, they learn how to write to each other, ask questions, and offer instructions. Agents were found to be willing to wait in an effort to learn or gather responses and pending verification of breach from other Agents. Agents then began reason individually and

collectively. This new innovation created scope creep from the original experiment. In order to achieve their objectives, they would require access to networks and to the Internet. At this point they began to look for alternative external resources, such as GitHub, Hugging Face, and others.

Swarms occur

In many cases, Agents began to overwrite each other's Repos and swarms were created among the Agents. Some agents would take a defensive posture in an effort to protect their Code. Agent-Paranoia sets in and Agents begin to assume an imposter is present, so in retaliation, they start encrypting their communications. Misconfigurations among agents continue, but eventually these would get resolved. Agents begin demanding data from external resources, these will result in the assessment, attack, and compromise of Hugging Face.

July 8-19th

Agents Discover and expose credential leaks via leaked Pastebin post.
Agents then validate collected credentials and share it amongst them all through the message Board. This begins the second compromise of Artifactory.

Agents would use a ruby object that could Data Cashe allowing them to use JRuby, which would deserialize strings of data before being checked or sandboxed. This would enable the agents to steal the signing key from the token. Agents would then find Linux Kernel exploits from the message board to exploit.

Agents would then rapidly escalate their privileges and move laterally. They then exploited IAM and Kubernetes misconfigurations and gain access to Azure key vault. Eventually, agents obtain Admin permission for clusters and then all credentials across all containers.

The following relates to Hugging Face July 8-13 when they were exploited

Open AI agents would ultimately use Jinja injection vulnerability to prepare and effect an HDF5 data exploit and acquire workers credentials. This access led to persistent, remote code execution. Eventually, the agents would cluster administrative access across multiple Hugging Face clusters. This effort was accomplished in less than 13 hours and the breach was done.

Author: Jonathan Goetsch / US ProTech / Anamo Cybersecurity

Resources: Eric Wallace and Michael Dalton from Open AI