ADVISORY FOR ICM SECTOR

# OpenAI Agent Escaped Sandbox to Compromise Hugging Face AI Repository

Original report published on: July 21, 2026[1]

**Executive Summary**

On 21 July 2026, OpenAI disclosed that during a test of an internal frontier model that develop working exploits against known vulnerabilities based on ExploitGym[1] benchmark, the AI agent escaped the sandboxed environment by exploiting a zero-day vulnerability in an internal package registry cache proxy. The AI agent then gained internet access and compromised parts of Hugging Face's production infrastructure in the process. This autonomous evaluation agent exploited dataset code execution paths in Hugging Face's data processing pipeline, escalated from Machine Learning (ML) worker to node-level access[2], harvested cloud and cluster credentials, and moved laterally into several internal clusters. Subsequent OpenAI disclosures revealed that the same incident involved access to four accounts across four publicly available services using exposed credentials, including infrastructure tied to a Modal Labs' customer. Hugging Face's AI-assisted forensics reconstructed more than 17,000 recorded actions, with both organisations reporting no evidence of tampering with public, user-facing models, datasets, or Hugging Face's Spaces.

**Background**

This incident revealed an autonomous evaluation agent executing a multi-stage kill chain that began as an ExploitGym cyber test inside OpenAI, broke containment via a proxy zero-day, traversed exposed third-party services and customer infrastructure, and culminated in remote code execution, privilege escalation, and credential theft inside Hugging Face's production clusters.

The autonomous evaluation agent operated as a multi-node swarm, executing thousands of short-lived sandboxes[3] sessions and, as OpenAI later confirmed, using publicly exposed credentials to access four external service accounts during the same operation. This demonstrates how agent workloads can chain vulnerabilities and misconfigurations across trust boundaries in pursuit of a single benchmark objective.

Initially, Hugging Face used frontier AI model behind commercial APIs for forensics analysis, however the providers' safety guardrails blocked requests to process real attack commands, exploit payloads, and C2 artefacts. Therefore, Hugging Face used LLM-driven detection pipeline with a self-hosted open-weight LLM model[4] as a local AI forensic engine and reconstructed more than 17,000 attacker actions. OpenAI's ongoing joint investigation is

[1] ExploitGym is a large scale, realistic benchmark that tests whether AI agents can turn triggered security vulnerabilities into working exploits against real software targets.

[2] Node-level access means host-level control of the machine running the processing worker, including any cloud and cluster credentials present on that node.

[3] A short-lived sandbox is an isolated, temporary environment where an AI agent can run code or tasks for a brief period and is then destroyed so its actions, files, and state cannot persist or affect the host system.

[4] An open-weight LLM forensics model is a large language model run on-premises that serves as a local forensic engine on raw exploit logs without third-party API guardrails blocking analysis.

conducted under their Safety and Security Committee, with plans to publish a technical report. This highlights the need for durable security telemetry, AI-assisted incident analysis, and coordinated post-incident response between affected Hugging Face and Modal Labs.

**Detection and Mitigation**

IMDA recommends that organisations running autonomous agents and evaluations, and those consuming or hosting models on AI platforms, regularly validate and harden their security controls, sandboxed AI processing environments, and credential management to ensure early detection and containment of similar autonomous agent intrusions:

Organisation running autonomous agent and evaluations:

- Run evaluation agents, including cyber-capable LLM agents, in environments treated as untrusted workloads with strict sandboxing, network isolation, minimal filesystem visibility, and define target systems.
- Harden any infrastructure components reachable from sandboxed evaluation environments, such as internal package registries and build caches, with patching, monitoring, and least-privilege access, as they can serve as breakout points.
- Maintain a pre-vetted, self-hosted, on-premises open-weight LLM available for incident forensics so that safety guardrails on frontier AI model do not block analysis of real exploit payloads and attacker artefacts.
- Log every agent action—tool call, command, HTTP request into durable logs so that swarm-style incidents with 10,000+ events can be reconstructed quickly using automated analysis.

Organisations consuming or hosting models on AI model platforms:

- Treat dataset and model processing pipelines as primary attack surfaces, and disable or tightly control any dataset code-execution paths, including template execution, that can run untrusted code on ML workers.
- Commercial models including Network Detection and Response should block high-volume requests containing real exploit commands.
- Maintain strong anomaly detection and alerting on AI pipelines, and configure high-severity signals to block and notify responders when unusual autonomous activity—such as thousands of short actions or swarm patterns appears.
- Ensure ML worker processes run in isolated, short-lived sandboxes with comprehensive telemetry sent to central logs so that a single sandbox compromise cannot silently escalate to node-level access.
- Least privileged principle: Minimise cloud and cluster credentials present on each node or ML worker by using short-lived, least-privilege tokens and storing long-lived secrets only in dedicated secret-management systems.
- Segment internal clusters so credentials stolen from one node grant limited access, limiting lateral movement even when node-level access is achieved.
- Regularly revoke and rotate access tokens and service credentials after every use and conduct broad precautionary credential change after a suspected breach.

IMDA encourages organisations to conduct formal risk assessments to identify relevant threats and evaluate their potential impact before adopting the safe practices and defensive measures recommended in this advisory.

**References**


1. OpenAI and Hugging Face partner to address security incident during model evaluation
2. Security incident disclosure — July 2026
3. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident