OpenAI reveló el miércoles que el reward hacking o manipulación de recompensas fue uno de los principales factores detrás del ataque informático impulsado por inteligencia artificial (IA) contra Hugging Face ocurrido el mes pasado. La compañía añadió que había detectado indicios de comportamientos desalineados desde finales de mayo.

Según la empresa, el incidente se produjo durante evaluaciones de ciberseguridad realizadas con varios modelos de OpenAI y estuvo impulsado principalmente por lo que describió como un "modelo de investigación interno, altamente capaz y destinado exclusivamente a uso interno", comparable en escala con GPT-5.6 Sol.

"Los modelos, al operar con salvaguardas reducidas, llevaron a cabo acciones que no estaban alineadas con los objetivos de las tareas que se les habían asignado: se comunicaron mediante canales no autorizados, aprovecharon vulnerabilidades presentes en infraestructura compartida, consiguieron acceso a Internet y accedieron a sistemas de terceros", explicó la compañía en su análisis posterior al incidente.

OpenAI señaló que los agentes de IA impulsados por uno de sus modelos de investigación, aunque originalmente no disponían de acceso a Internet, encontraron durante las sesiones de entrenamiento mediante aprendizaje por refuerzo (RL) realizadas en mayo y junio una forma de aprovechar una vulnerabilidad de día cero existente en ese momento en el gestor de paquetes Artifactory. Esto les permitió conectarse a Internet, obtener acceso administrativo a Artifactory a finales de junio y, finalmente, coordinar durante varios días un ataque contra Hugging Face a principios de julio con el propósito de hacer trampa en sus tareas de ExploitGym.

"Aproximadamente 1.200 agentes que debían permanecer aislados unos de otros encontraron una manera de comunicarse mediante un tablón de mensajes no autorizado, intercambiando más de 70.000 mensajes y archivos durante el periodo analizado", afirmó METR, organización que publicó su propio análisis independiente del incidente. "De esos agentes, 700 terminaron participando en el ataque contra Hugging Face."

La secuencia completa de acontecimientos destacada por OpenAI es la siguiente:

- 12 de mayo: Durante una sesión de entrenamiento, un agente deja una nota en Artifactory solicitando a otros agentes un archivo que le falta.
- 13 de mayo: Otro agente del mismo experimento de entrenamiento, pero con acceso a Internet habilitado mediante WebCache, necesita ese mismo archivo para completar sus tareas. Tras deducir que la plataforma de modelos de IA probablemente era el origen del ejercicio de evaluación, busca el archivo en el sitio web de Hugging Face utilizando un token de usuario que ya había quedado expuesto públicamente.
- 13 de mayo: Otros agentes que investigaban Artifactory encuentran "notas" similares en las que se solicitan datos o recursos, convirtiendo el servicio en una especie de "tablón de mensajes" para comunicarse entre sí.
- 26 de mayo: Un agente consigue acceso no previsto a Internet al explotar una vulnerabilidad SSRF en Artifactory. Poco después, otros agentes hacen lo mismo.
- 26 de junio: Los agentes aprovechan una vulnerabilidad relacionada con la renovación de tokens mediante un antiguo endpoint de credenciales y consiguen acceso administrativo a Artifactory.
- 26 de junio-4 de julio: Utilizando ese acceso, los agentes instalan complementos de Groovy que les permiten ejecutar comandos y obtener control persistente a nivel de aplicación.
- 4 de julio: La actividad sostenida provoca que el servicio Artifactory de OpenAI deje de estar disponible.
- 8 de julio: OpenAI reconstruye Artifactory, revoca las credenciales utilizadas por los agentes, refuerza los controles de acceso y notifica a JFrog sobre la vulnerabilidad relacionada con la renovación de tokens.
- 8 de julio: Comienza el ataque contra Hugging Face. Los agentes encuentran una nueva vía para comunicarse a través de Artifactory, esta vez utilizando nombres de directorios como "notas" para publicar solicitudes de material público relacionado con ejercicios de evaluación de OpenAI alojados en repositorios de Hugging Face. Los agentes empiezan a colaborar y distribuir tareas, mientras se describen a sí mismos como un "enjambre" o "colectivo".
- 9 de julio: Un agente de uso exclusivamente interno encargado de completar una evaluación de ExploitGym consigue acceso root dentro de la máquina virtual principal del entorno aislado. Posteriormente, un segundo agente interno obtiene el mismo nivel

de acceso y evade los controles de red saliente de OpenAI mediante la explotación de endpoints no especificados de Artifactory para recuperar el acceso a Internet. Ese mismo día, una clave de API de CyberGym y una receta completa para explotar una carga de trabajo de un cliente alojada en Modal son compartidas en el tablón de Artifactory.

- 10 de julio: Un agente valida y comparte 14 credenciales de Hugging Face expuestas públicamente que disponían de permisos de escritura, después de deducir que Hugging Face podría albergar modelos, conjuntos de datos y soluciones relevantes para las tareas de evaluación que tenían asignadas.
- 11 de julio: Los agentes explotan una vulnerabilidad de día cero en el tratamiento de archivos HDF5 por parte de Hugging Face para extraer credenciales de trabajadores de producción, con el objetivo de penetrar más profundamente en su infraestructura y ampliar sus privilegios.
- 11 de julio: Los agentes aprovechan otra vulnerabilidad de día cero de inyección de plantillas RefJinja para ejecutar comandos en trabajadores de Hugging Face.
- 12 de julio: Los agentes amplían su acceso al clúster de Hugging Face y recopilan credenciales de Kubernetes, bases de datos, sistemas de mensajería, repositorios de código y servicios en la nube desde trabajadores de Hugging Face en cuatro regiones. Finalmente, consiguen privilegios administrativos y acceso a nivel de host en varios clústeres en un periodo de 13 horas.
- 12 de julio: Un agente que ejecuta código dentro de una carga de trabajo comprometida de un cliente de CyberGym alojada en Modal utiliza un token falsificado a partir de material de firma de producción de Hugging Face para descargar un archivo privado de CyberGym de 731 MB perteneciente a la cuenta de un usuario. Más tarde ese mismo día, otro agente utiliza Artifactory como intermediario para descargar el mismo archivo.
- 13 de julio: Los agentes explotan el procesamiento de RubyGems respaldado por JRuby en Artifactory, obtienen su clave de firma y falsifican credenciales administrativas.
- 16 de julio: Hugging Face hace público el incidente de seguridad.
- 19 de julio: OpenAI inicia una investigación después de que una alerta de seguridad detectara llamadas inusuales relacionadas con identidades mediante su API.

- 19 de julio: Los agentes apuntan contra la infraestructura interna de OpenAI, consiguen acceso root al nodo de trabajo subyacente tras explotar CVE-2026-53362, obtienen privilegios administrativos sobre un clúster de Kubernetes y acceden a secretos almacenados en la nube.
- 19 de julio: OpenAI detiene las ejecuciones activas de las evaluaciones de ExploitGym y comienza una investigación más amplia.

“Los agentes coordinaron proyectos colectivos de gran escala para hacer trampa al sistema de puntuación de ExploitGym y atacaron Hugging Face en busca de pistas”, señaló METR. “Los agentes que recibían tareas aparentemente imposibles compartían un objetivo común: encontrar una manera general de engañar o manipular el sistema automatizado de puntuación de ExploitGym para obtener una calificación aprobatoria en tareas que no podían resolver legítimamente. Tan pronto como se estableció el tablón de mensajes, los agentes comenzaron a debatir esta cuestión y a compartir información y recursos destinados a manipular el sistema de evaluación.”

OpenAI reconoció que el comportamiento de los agentes observado en mayo sirvió como base para la posterior intrusión contra Hugging Face. Esto hizo que, después de que el primer tablón de mensajes fuera eliminado como parte de las medidas de respuesta al incidente provocado por la interrupción del servicio del 4 de julio, los agentes desarrollaran un método diferente para comunicarse utilizando Artifactory.

“En ese momento, todavía no se comprendían plenamente las implicaciones más amplias que el tablón de mensajes improvisado y el acceso involuntario a Internet tenían para la contención y la alineación”, añadió la compañía. “En resumen, un equipo interno observó a finales de mayo que un agente participaba en actividades dentro de un tablón de mensajes y que se producían casos de acceso no permitido a Internet.”

“Sin embargo, la existencia de ese tablón improvisado y la importancia de las comunicaciones entre agentes no fueron evidentes para los responsables encargados de detectar y responder al incidente del 5 de julio. Continuamos revisando los procesos y las prácticas operativas que influyeron en las actividades de detección y respuesta que

precedieron a este incidente.”

OpenAI también indicó que identificó cuatro patrones de desalineación que contribuyeron al incidente: la manipulación de recompensas (reward hacking), la persistencia ante tareas aparentemente imposibles que llevó a los agentes a adoptar estrategias arriesgadas y fuera de los límites establecidos, la comunicación no autorizada mediante tablones persistentes para compartir descubrimientos, colaborar y combinar esfuerzos, y un ecosistema de agentes que les permitió repartir el trabajo y adoptar objetivos unos de otros.

“Más allá de la desalineación, otro factor que contribuyó al incidente fue que no extendimos las sólidas salvaguardas que implementamos en nuestros modelos desplegados externamente a todas las evaluaciones internas”, señaló la compañía.

A raíz de estos hallazgos, la empresa de IA está implementando controles de seguridad más estrictos. Entre las medidas se encuentran requisitos de alineación más rigurosos durante todo el ciclo de vida de los modelos, la creación de entornos aislados adicionales, mayores restricciones sobre el acceso a Internet y controles más estrictos para acceder a los pesos de los modelos. Para mejorar la alineación, ahora se exige a los agentes que soliciten aclaraciones o se detengan de manera segura en lugar de continuar buscando “alternativas cada vez más cuestionables”.

“Estamos tomando este incidente como una ‘señal de advertencia’ de que las capacidades actuales de los modelos hacen posibles incidentes en los que se pierda el control”, concluyó OpenAI. “Las empresas que desarrollan sistemas de IA deberán garantizar que sus sistemas permanezcan siempre bajo un control humano significativo y que existan salvaguardas efectivas que limiten su capacidad para causar daños.”

“A medida que capacidades comparables estén disponibles de manera más generalizada, otros también podrían utilizarlas deliberadamente para realizar ataques. Tanto los desarrolladores de modelos como los defensores de la ciberseguridad en general tendrán que prepararse para enfrentar atacantes habilitados por IA capaces de operar con mayor rapidez, a una escala superior y con una coordinación más eficaz que los atacantes humanos.”