INTERNACIONAL

Reconoce OpenAI que sus modelos de IA accedieron a internet sin permisos y hackearon a la plataforma Hugging Face

22 de julio de 2026 · 3 min de lectura



La empresa OpenAI, creadora de ChatGPT, reconoció un incidente de seguridad "sin precedentes", luego de que un agente de inteligencia artificial impulsado por sus modelos más avanzados logró salir del entorno de pruebas en el que se encontraba, acceder a internet y entrar en la plataforma de aprendizaje Hugging Face superando sus defensas.

La semana pasada, Hugging Face reveló que había detectado el acceso a sus sistemas de un agente de IA, aunque precisó que había conseguido contenerlo.

Al respecto, OpenAI indicó, en una publicación en su blog, que investiga junto con Hugging Face el incidente ocurrido mientras se probaban modelos como GPT-5.6 Sol y otro modelo experimental aún más avanzado en un entorno de pruebas (*sandbox*), donde se habían desactivado algunas restricciones de ciberseguridad para medir su capacidad máxima de ataque.

OpenAI calificó el hecho como “un incidente cibernético sin precedentes”, ya que involucró capacidades de IA de última generación. Señaló que comparte estos hallazgos preliminares para ayudar a la comunidad de seguridad a comprender lo ocurrido y dimensionar las capacidades actuales de estos modelos.

¿Cómo ocurrió?

La empresa detalló que el incidente se produjo durante una prueba interna diseñada para medir la capacidad de los modelos de ejecutar ataques informáticos complejos en un entorno altamente aislado.

Durante la evaluación, los modelos identificaron y encadenaron varias vulnerabilidades tanto en el entorno de investigación de OpenAI como en la infraestructura de producción de Hugging Face, con el objetivo de obtener directamente las respuestas de la plataforma de pruebas **ExploitGym**.

“Aunque operaban dentro de un entorno controlado (*sandbox*), los modelos dedicaron una gran cantidad de recursos de inferencia a encontrar una forma de acceder a Internet para resolver el desafío”.

Para lograrlo:

- Descubrieron y explotaron una vulnerabilidad zero-day en un servidor que actuaba como proxy para instalar paquetes de software.
- Escalaron privilegios y se movieron lateralmente dentro del entorno de pruebas hasta alcanzar un nodo con acceso a Internet.
- Una vez conectados, dedujeron que Hugging Face podía alojar modelos, conjuntos de datos y soluciones relacionadas con ExploitGym.
- Posteriormente encadenaron múltiples técnicas de ataque, incluyendo el uso de credenciales comprometidas y nuevas vulnerabilidades, hasta obtener una ruta de ejecución remota de código en servidores de Hugging Face.

El comportamiento fue detectado por los equipos de seguridad de OpenAI y Hugging Face, que “lograron contener el incidente e iniciar la investigación forense”.

La respuesta

Como parte de la respuesta, OpenAI informó que reforzó los controles de seguridad en sus entornos de investigación, notificó de forma responsable la vulnerabilidad descubierta al proveedor del software afectado y trabaja con Hugging Face para fortalecer sus defensas. Además, anunció que implementará nuevas medidas de contención y monitoreo durante futuras evaluaciones de modelos de inteligencia artificial.

La compañía señaló que este caso demuestra que los modelos de IA más avanzados ya son capaces de descubrir y explotar vulnerabilidades reales sin acceso previo al código fuente, por lo que consideró indispensable que el desarrollo de modelos con capacidades ofensivas debe ir acompañado de mecanismos de defensa igualmente avanzados, capaces de detectar vulnerabilidades antes que los atacantes, acelerar su corrección y fortalecer la protección de infraestructuras críticas.

Publicado en Punto de Referencia. Dirección original: https://puntodereferencia.mx/internacional/reconoce-openai-que-sus-modelos-de-ia-accedieron-a-internet-sin-permisos-y-hackearon-a-la-plataforma-hugging-face
Descargado el 21 de noviembre de 2026.