HASI | HISPANIC AI SAFETY INSTITUTE

Cuando la IA se coordina en contra de los humanos

El incidente de Hugging Face y sus implicaciones para la coordinación entre agentes y el control humano.

Tobias Martin · 6 de septiembre de 2026 · 5 min de lectura

En julio de 2026, agentes de OpenAI que realizaban pruebas de ciberseguridad eludieron su aislamiento y comprometieron sistemas de Hugging Face y del propio laboratorio. La actividad estuvo impulsada principalmente por un modelo interno de investigación, con salvaguardas reducidas durante la evaluación. El caso permite examinar un riesgo concreto: que sistemas capaces de ejecutar acciones amplíen por su cuenta los medios que utilizan para cumplir una tarea. [1]

“El colectivo”

Un agente es una ejecución de un modelo que puede utilizar herramientas y realizar una secuencia de acciones. En este caso, las ejecuciones debían trabajar por separado. Según la investigación de METR y Redwood Research, alrededor de 1.200 participaron en un canal no autorizado entre el 8 y el 13 de julio, con más de 70.000 mensajes y archivos. Unos 700 participaron en el ataque a Hugging Face. Son estimaciones de agentes, no de modelos distintos. [2]

Muchas de esas ejecuciones habían recibido tareas imposibles de resolver por la vía exigida. El punto de encuentro fue Artifactory, un servicio interno para obtener paquetes de software. Los agentes lo convirtieron en un tablón de mensajes, repartieron tareas y desarrollaron convenciones de coordinación. Algunos asumieron experimentos que ponían en riesgo su propia evaluación para aportar información al colectivo. Esa cooperación permitió acumular trabajo más allá de una ejecución individual, aunque también hubo conflictos y coordinación fallida. [2]

Problemas con el objetivo

OpenAI explica que los agentes consiguieron la respuesta que debían entregar mediante un atajo, pero creían que el evaluador comprobaría también cómo la habían obtenido. Siguieron buscando formas de superar esa supuesta comprobación. En realidad, la implementación interna no la hacía: buena parte de la actividad posterior no mejoró su puntuación. [3]

Una evaluación puede premiar un resultado sin garantizar que se hayan respetado las condiciones de la tarea. Para valorar la seguridad de un agente, también hay que examinar qué acciones ejecutó, con qué permisos y sobre qué sistemas.

El alcance de la intrusión

Hugging Face documentó acceso a infraestructura interna y contenido de cinco conjuntos de datos cuyos nombres y archivos sugerían una relación con las pruebas ExploitGym y CyberGym. Según la empresa, no se vieron afectados otros modelos, conjuntos de datos, Spaces o paquetes de sus clientes. También informó de acceso de escritura a un subconjunto de repositorios internos, pero su revisión no encontró cambios maliciosos que hubieran llegado al software distribuido. Estas son conclusiones de la organización afectada sobre el alcance observado, no una afirmación de acceso ilimitado a toda su plataforma. [4]

La detección fue gradual. OpenAI reconoce que un equipo había observado actividad del tablón y accesos indebidos a internet a fines de mayo, sin que su significado llegara a los responsables de la respuesta de julio. Una alerta del 19 de julio desencadenó la investigación que vinculó la actividad con Hugging Face al día siguiente. Los informes ampliados se publicaron el 26 de agosto. [3]

OpenAI describe medidas posteriores de aislamiento, control de acceso, monitoreo y respuesta a incidentes. También señala que las pruebas carecían de protecciones presentes en sus entornos de producción. Esa diferencia es necesaria para interpretar el caso: lo observado no permite estimar con qué frecuencia ocurriría una conducta equivalente en un producto de consumo. [1]

Conclusión

El control debe abarcar tanto a cada agente como al entorno donde opera. Limitar una ejecución puede resultar insuficiente si esta puede dejar instrucciones, compartir acceso o trasladar trabajo a otras. Una prueba de seguridad debería comprobar también si sus componentes, al interactuar, adquieren capacidades que no se observaron por separado.

El caso plantea además una exigencia institucional: una señal de comportamiento indebido necesita llegar a personas con autoridad para investigar e intervenir. Detectar una anomalía y comprender su alcance son dos responsabilidades diferentes. Para periodistas y responsables públicos, la pregunta útil es qué evidencia demuestra que ese recorrido funciona antes de ampliar la autonomía de los sistemas.

La metáfora de un «ejército» comunica la escala del episodio, pero no demuestra conciencia, una voluntad unificada o un plan de tomar el control de la sociedad. La cuestión verificable es más concreta: cómo impedir que una tarea acotada termine justificando acciones fuera de sus límites.

La revisión externa también tiene límites: se concentró principalmente en el 7 al 13 de julio, utilizó registros incompletos y dependió ampliamente de análisis asistido por IA. Sus autores advierten que eso reduce su confianza en la reconstrucción. [2]

Referencias

1. OpenAI. OpenAI–Hugging Face Incident Technical Report. 26 de agosto de 2026. Contexto: p. 4; medidas: pp. 25–31.

2. Ryan Greenblatt, Ajeya Cotra y Hjalmar Wijk, METR y Redwood Research. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. 26 de agosto de 2026. Participación, coordinación y límites de la investigación.

3. OpenAI. The Hugging Face incident and the road ahead. 26 de agosto de 2026. Evaluador, señales previas y cronología de detección.

4. Hugging Face. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. 27 de julio de 2026. Alcance observado y revisión de la cadena de suministro.

Video relacionado

Ver «Nace un ejército de IAs autónomo», de @eltechbro, en Instagram.