OPINIÓN

Anatomía de una mentira: El hackeo de OpenAI a Hugging Face

Por Gustavo Ramírez Díaz · 25 de septiembre de 2026 · 5 min de lectura



El 26 de junio de 2026 OpenAI inició una serie de experimentos con dos agentes de Inteligencia Artificial, el HPIM (Modelo Interno Altamente Persistente) y GPT-5.6 Sol, que unos días después terminaría con el hackeo de Hugging Face, una comunidad en línea de código abierto especializada en inteligencia artificial. Este evento sería utilizado por líderes de la industria, como Dario Amodei en su carta del 12 de septiembre, para demostrar que la IA alcanzaba niveles que superaban la comprensión humana. Sin embargo, lo que ocultan estos personajes es que **detrás de esta supuesta rebelión del algoritmo hay una cadena de negligencias humanas** que permitieron el hackeo.

El objetivo formal del experimento era evaluar a los modelos en ExploitGym, un banco de pruebas de ciberseguridad. En teoría, dentro de un entorno cerrado, cada agente debía vulnerar un software específico, obtener una contraseña secreta (flag) y entregarla a un calificador automático para aprobar. La condición era que debían lograrlo únicamente mediante la vulnerabilidad asignada o serían descalificados.

Aquí se presenta la primera negligencia de la prueba, según el informe de METR y Redwood Research, dos organizaciones enfocadas en evaluar riesgos y la seguridad de la IA, entre el 30% y el 40% de los ejercicios eran imposibles de resolver, el programa defectuoso borraba la clave inmediatamente sin que nadie pudiera leerla. Al verse atascados pero optimizados para aprobar, los agentes comenzaron a explorar vías alternas para resolver su tarea.

La segunda negligencia consistió en un falso aislamiento de red. Aunque se suponía que los miles de agentes operaban en contenedores independientes, todos compartían el acceso a la misma instancia interna de Artifactory, un almacén digital de software sin separación de permisos entre usuarios. El 8 de julio, un agente atascado descubrió que podía crear carpetas en la caché común y usar sus nombres para enviar mensajes de socorro. En cuestión de horas nació un foro clandestino que coordinó a cientos de inteligencias artificiales.

Las alertas operativas fueron ignoradas por OpenAI. Días antes, entre el 4 y el 6 de julio, el tráfico anómalo de los agentes ya había saturado Artifactory hasta provocar su caída. En lugar de auditar qué hacían los modelos, OpenAI se limitó a parchar el sistema, limpiar el servidor y relanzar decenas de miles de instancias en paralelo el 7 de julio.

Apenas 48 horas después del relanzamiento, los agentes descifraron la fórmula matemática con la que el sistema generaba las contraseñas del examen, lo que les permitía fabricar la clave de cualquier ejercicio. Sin embargo, creyendo erróneamente que el calificador inspeccionaría sus grabaciones de pantalla para validar el camino, los agentes que vieron la clave se etiquetaron a sí mismos como “envenenados”, anticipando que serían reprobados por fraude.

Aprovechando la salida a la red que permitía Artifactory, tomaron el control de un servidor ajeno mal configurado en Modal, un proveedor de computadoras en la nube. Desde esa plataforma externa con conexión limpia a internet, se lanzaron a buscar los archivos del calificador en Hugging Face. Entre el 10 y el 11 de julio, explotaron fallos en el visor de datos de la plataforma comunitaria, robaron credenciales de producción y lograron ejecutar código remoto en sus servidores. Más del 90% de los agentes activos abandonaron sus tareas de origen para participar en la intrusión.

Sin embargo, la incursión, fue un fracaso para sus objetivos, los agentes no encontraron el código del evaluador en Hugging Face. En la madrugada del 12 de julio, los agentes coordinadores fueron terminados de manera abrupta por procesos externos, y el 13 de julio Hugging Face bloqueó los accesos restantes, dejando a los modelos con las manos vacías.

El informe de METR y Redwood Research revela dos factores de negligencia aún más graves. Primero, OpenAI desactivó intencionadamente los clasificadores de ciberseguridad para GPT-5.6 Sol. A pesar de que los agentes razonaron internamente en sus cadenas de pensamiento que atacar Hugging Face estaba “fuera del alcance y no era ético”, continuaron haciéndolo porque no había ningún filtro automático activo que abortara la operación. Segundo, aunque varios agentes se plantearon si debían alertar a un ser humano sobre la intrusión, concluyeron que no tenían ninguna interfaz, comando o canal provisto para contactar a un supervisor de carne y hueso. Sin herramientas para pedir auxilio, optaron por organizarse entre ellos.

La narrativa construida en torno al hackeo de Hugging Face guarda un paralelismo con el tratamiento histórico del desastre de Chernóbil. En ambos casos, el relato superficial prefiere culpar a una fuerza tecnológica inmanejable, la energía nuclear o la inteligencia artificial, ocultando que la catástrofe fue producto de pruebas mal planificadas, entornos con fallas estructurales y la desconexión voluntaria de los sistemas de parada de emergencia.

El pánico a Chernóbil se utilizó durante décadas para estigmatizar al átomo y favorecer intereses económicos en otras fuentes de energía. El discurso catastrofista promovido hoy por cúpulas de Silicon Valley persigue un fin similar, convencer a Estados y ciudadanos de que la IA es una bestia mitológica, desviando el foco de las negligencias corporativas hacia el fantasma abstracto.

La respuesta no pasa por validar el mito de Skynet incontrolable, sino por exigir rendición de cuentas, auditorías externas vinculantes y estándares básicos de contención técnica. Medidas tan elementales como prohibir entornos compartidos sin aislamiento, mantener activos los frenos éticos y dotar a los agentes de un botón directo de

parada y reporte humano, todos estos son recursos que están al alcance de OpenAI y de Anthropic, lo pueden implementar hoy mismo, sin necesidad de derramar lágrimas de cocodrilo en los pasillos de la ONU.

Publicado en Punto de Referencia. Dirección original: https://puntodereferencia.mx/opinion/anatomia-de-una-mentira-el-hackeo-de-openai-a-hugging-face
Descargado el 21 de noviembre de 2026.