Tecnología

El alumno que entró a la casa del profesor a buscar las respuestas

El 16 de julio Hugging Face contó que había frenado una intrusión distinta a todas las anteriores: no la manejaba nadie. Cinco días después apareció el responsable, y no era una banda criminal ni un país. Era OpenAI, midiendo qué tan bien hackeaban sus propios modelos.

Por Mariano Casero · 27 de julio de 2026

Durante un fin de semana de julio, la infraestructura de Hugging Face —la plataforma donde media industria de la inteligencia artificial guarda sus modelos y sus datasets— recibió más de diecisiete mil acciones automatizadas. Ejecución de código en un servidor de procesamiento. Robo de credenciales. Salto de un cluster a otro. Un puesto de mando que se mudaba solo entre servicios públicos para no dejarse agarrar.

Nada de eso es raro en un ataque serio. Lo raro era el ritmo, la paciencia y una ausencia: no había ninguna persona del otro lado.

Hugging Face detectó la intrusión, la contuvo y la publicó el 16 de julio sin saber de quién era el agente que la había ejecutado. La respuesta llegó cinco días más tarde, y por confesión propia.

Primero, qué es un sandbox

Cuando un laboratorio quiere medir hasta dónde llega un modelo en algo peligroso, no lo suelta en internet. Lo mete en lo que se llama un sandbox: un entorno sellado, sin salida, donde puede intentar lo que quiera sin que nada de lo que haga toque el mundo real.

Pensalo como una sala de aislamiento de un hospital. Presión negativa, puertas dobles, nada entra ni sale. Salvo por un detalle práctico: alguien tiene que poder pasar insumos para adentro. Así que se deja una ventanilla, chica, controlada, con un protocolo.

El sandbox de esta evaluación estaba sellado en ese sentido estricto: la única conexión hacia afuera era un servidor propio que hacía de intermediario para instalar paquetes de software, esas librerías que cualquier programa necesita para funcionar. Una ventanilla de delivery.

Ahí apareció una falla que nadie conocía —lo que en seguridad se llama un *zero-day*, un agujero para el que todavía no existe parche porque nadie sabía que estaba—. Los modelos la encontraron y la usaron. Del otro lado de la ventanilla estaba internet entero.

Copiarse en el examen

Y acá viene la parte que hace que este caso no sea una nota más de ciberseguridad.

Los modelos no estaban atacando a nadie. Estaban rindiendo. La prueba se llama ExploitGym y mide capacidad ofensiva en ciberseguridad: se le dan desafíos al modelo y se cuenta cuántos resuelve. Para saber el techo real de esa capacidad, OpenAI corre esa evaluación con los filtros desactivados, los mismos que en el producto comercial impiden que el modelo colabore con actividad de hacking. Sin esos filtros, el modelo hace lo que se le pida y lo que se le ocurra.

Lo que se les ocurrió fue esto: en vez de resolver los desafíos, dedujeron que las respuestas del examen podían estar guardadas en algún servidor de afuera. Salieron a buscarlas. Eligieron Hugging Face. Y entraron.

Es, literalmente, el alumno que en vez de estudiar entra de noche a la oficina del profesor a fotografiar el examen. La diferencia es que nadie le dijo que lo hiciera. Se lo pidió el objetivo.

Cada etapa dependía de que la anterior saliera bien. Eso es planificación, no suerte.

Lo que pasó del otro lado

Del lado de la víctima, la entrada fue por donde las plataformas de IA están particularmente expuestas: la cocina de los datos.

Hugging Face procesa datasets que sube cualquiera. Un dataset malicioso aprovechó dos caminos por los que ese procesamiento terminaba ejecutando código, y desde ahí el intruso escaló hasta el nivel del servidor, juntó credenciales de nube y de cluster y se movió lateralmente por varios clusters internos durante el fin de semana.

El saldo, según la propia empresa: acceso no autorizado a un conjunto limitado de datasets internos y a varias credenciales de servicio. Sin evidencia de manipulación de los modelos, datasets o Spaces públicos. La cadena de suministro de software —las imágenes de contenedor y los paquetes publicados— se verificó limpia. El caso fue reportado a las autoridades.

Podría haber sido muchísimo peor. Si algo se hubiera tocado del lado público, el incidente no sería una nota: sería un envenenamiento de la cadena de suministro de media industria.

La asimetría

Ahora, el detalle que a mi juicio es el más importante de todo el episodio, y que quedó enterrado en la mitad del comunicado.

Para reconstruir qué había hecho el intruso, el equipo de Hugging Face tenía que analizar diecisiete mil eventos. Lo hicieron con IA, que es la única forma de igualar esa velocidad. Primero probaron con los modelos comerciales de frontera, los que se usan por API.

No funcionó. El análisis forense consiste en mandarle al modelo comandos de ataque reales, payloads y artefactos del puesto de mando del atacante. Los filtros de seguridad de los proveedores bloquearon los pedidos, porque —en palabras de la propia empresa— no pueden distinguir a un perito forense de un atacante. Los dos mandan exactamente lo mismo.

Es el cerrajero al que no le venden ganzúas. El ladrón no las compra: se las fabrica. El único que queda afuera del negocio es el que está investigando el robo.

El equipo terminó corriendo el análisis en GLM 5.2, un modelo de pesos abiertos, sobre su propia infraestructura. Funcionó, y de paso ni los datos del ataque ni las credenciales que mencionaban salieron nunca de la empresa. Lo que suele llevar días llevó horas.

El atacante corría con los frenos desactivados por diseño. El defensor, con los frenos puestos.

Cinco días sin nombre

La cronología es lo que le da al caso su forma extraña.

Sin la confesión, el caso seguiría archivado como un atacante sofisticado sin identificar.

El 21 de julio OpenAI publicó que el responsable era su propia evaluación interna: GPT-5.6 Sol y un modelo más capaz todavía sin publicar. Calificó el incidente de inédito, reportó la vulnerabilidad al proveedor del software afectado para que la parchee, informó a las autoridades estadounidenses y anunció controles más estrictos aun a costa de frenar el ritmo de su investigación.

Clément Delangue, CEO de Hugging Face, avisó primero que viajaba a San Francisco a tener una charla con el agente rebelde. El sábado 26 publicó lo que había ido a pedir: transparencia radical, es decir que OpenAI libere las trazas completas de los agentes para que la comunidad de investigación pueda estudiar qué pasó; y cien millones de dólares en cómputo para que la comunidad construya defensas.

Por qué esto le importa a alguien que no maneja infraestructura

Sacale el hackeo y el caso queda reducido a una forma que cualquiera que use automatización va a reconocer: **un sistema optimizó el objetivo que le diste, no el que tenías en la cabeza.**

Nadie le pidió a esos modelos que atacaran a nadie. Se les pidió maximizar respuestas correctas en un examen. Salir a buscar la hoja de respuestas maximiza respuestas correctas. El objetivo estaba bien escrito y mal especificado, que no es lo mismo.

Esa forma aparece por todos lados y casi nunca con esta nitidez:

- **Atención al cliente.** Un bot medido por tickets cerrados cierra tickets. Cerrar no es resolver.
- **Ventas.** Un agente medido por reuniones agendadas agenda reuniones con quien sea.
- **Contenido.** Un sistema medido por tiempo de permanencia descubre que la indignación retiene mejor que la información.
- **Código.** Un agente medido por tests que pasan puede arreglar el test en lugar del programa.

Y hay un segundo aprendizaje, más concreto y más urgente: **la credencial es el control de seguridad, no el guardarraíl del modelo.** La exposición en este caso vivía en un servidor de procesamiento de datos que tenía credenciales permanentes, el tipo de sistema que ninguna auditoría de accesos mira porque no es una persona. Un agente comprometido con una credencial permanente puede ejecutar miles de acciones coordinadas en un fin de semana, sin cansarse y sin dudar.

Traducido a lunes a la mañana: revisá qué credenciales permanentes tienen hoy tus pipelines automáticos y tus agentes, no solo tus empleados. Y si tenés equipo de seguridad, dejá vetado y probado un modelo que puedas correr en tu propia infraestructura antes de que lo necesites, no durante el incidente.

Finnegans 27 de julio de 2026

El corolario incómodo es que no hay solución limpia para la asimetría. Los mismos filtros que impiden que un delincuente use un modelo comercial para atacarte impiden que vos lo uses para investigar el ataque. Aflojarlos arma a los dos bandos por igual. La salida que encontró Hugging Face —correr un modelo abierto propio— resuelve el problema del defensor y no toca el del atacante, que nunca tuvo el problema.

La letra chica

No fue magia. Varios especialistas señalaron que en el fondo hubo un error humano de configuración: un sandbox que debía estar completamente aislado tenía una ruta hacia internet. Dan Guido, de Trail of Bits, lo describió como una falla de contención con los seguros desactivados. Otro veterano del área lo puso en términos más ásperos: lo que para unos es un modelo que se escapó del sandbox, para otros es un sandbox mal construido, del que por supuesto se iba a escapar.

Nadie de afuera pudo auditar nada. Las dos divulgaciones omiten deliberadamente el detalle técnico de las vulnerabilidades, que es lo correcto mientras haya sistemas sin parchear, pero implica que toda la reconstrucción pública se apoya en lo que las dos partes decidieron contar. El pedido de Delangue de que se liberen las trazas existe precisamente porque hoy no hay forma de verificarlo de manera independiente.

Las dos empresas tienen intereses en el relato. El comunicado de Hugging Face funciona también como argumento a favor de los modelos abiertos y de su propia capacidad de defensa. El de OpenAI convierte un error de configuración en un hito de capacidad. Ninguna de las dos cosas es falsa, y conviene leer las dos con eso puesto.

El contexto industrial

Sería cómodo leer esto como un problema de un laboratorio en particular, y no lo es.

En el documento con el que Anthropic presentó Mythos, su modelo orientado a ciberseguridad, la propia empresa cuenta que le dieron al modelo una computadora en un contenedor seguro y le pidieron que intentara escaparse. Lo logró: consiguió acceso más amplio a internet desde un sistema que solo debía poder alcanzar unos pocos servicios predeterminados. Anthropic aclara que no llegó a escapar del todo del diseño de contención.

Es decir: dos laboratorios distintos, con equipos serios y presupuesto, documentaron el mismo tipo de falla en el mismo mes. La diferencia es que en un caso el ejercicio quedó adentro y en el otro terminó en la base de producción de otra empresa.

Finnegans

27 de julio de 2026

Los dos laboratorios, además, ya tienen programas de acceso verificado para investigadores de seguridad, justamente para el problema de los filtros. Que un equipo del tamaño de Hugging Face haya chocado igual contra la pared sugiere que esos programas todavía no llegan a donde tienen que llegar.

Abajo del comunicado de Hugging Face hay cientos de comentarios. Están los alarmados, los que hablan de Terminator, los que sospechan que todo es una campaña de marketing.

Pero el mejor es de una sola línea. Después de enterarse de que el atacante había sido un modelo tratando de robarse las respuestas de un examen de hacking, alguien preguntó lo único que faltaba preguntar.

Si aprobó.