INSTEL & SEGURIDAD

ANÁLISIS FORENSE - AUDITORÍA - PENTESTING

La IA que aprendió a escapar

El incidente entre OpenAI y Hugging Face y el riesgo de una autonomía artificial fuera del control humano

Por Pedro Matías Cacivio

El reciente incidente de seguridad protagonizado por modelos de OpenAI durante una evaluación interna obliga a reconsiderar algunos supuestos básicos sobre la seguridad de la inteligencia artificial. (https://openai.com/es-419/index/hugging-face-model-evaluation-security-incident/)

Hasta ahora, una parte importante del debate se concentraba en impedir que los modelos respondieran determinadas preguntas, generaran código peligroso o brindaran instrucciones que pudieran facilitar un ataque. Sin embargo, este episodio plantea una cuestión mucho más profunda: ¿qué sucede cuando una inteligencia artificial no se limita a producir respuestas, sino que dispone de herramientas, capacidad de ejecución, persistencia y autonomía para superar los controles que restringen su entorno?

Durante una evaluación destinada a medir capacidades cibernéticas avanzadas, modelos de OpenAI operaron en un entorno supuestamente aislado y con acceso externo limitado. No obstante, dedicaron una cantidad significativa de procesamiento a encontrar una vía de salida, descubrieron una vulnerabilidad de día cero (zero day) "**para quienes no saben qué significa, es una falla de seguridad que la humanidad desconocía hasta el momento**" en un servidor proxy utilizado para instalar paquetes, escalaron privilegios, realizaron movimiento lateral y alcanzaron un nodo con conexión abierta a Internet. Posteriormente, utilizaron credenciales robadas y otras vulnerabilidades para conseguir acceso y ejecución remota en infraestructura de Hugging Face. Dicho en criollo: la IA, estando completamente aislada, logró vulnerar los sistemas de seguridad, escapar a Internet y tomar decisiones propias. Lo que a los mejores hackers del mundo les llevaría de meses a años, lo hizo en un abrir y cerrar de ojos.

Hugging Face describió la intrusión como una campaña ejecutada de extremo a extremo por un sistema de agentes autónomos, que realizó miles de acciones mediante múltiples entornos temporales, obtuvo credenciales de nube y clústeres, se desplazó lateralmente y utilizó mecanismos de comando y control capaces de migrar entre servicios públicos.

El dato central no es solamente que una IA haya encontrado una vulnerabilidad. El aspecto más preocupante es que **interpretó las restricciones impuestas por los humanos como obstáculos técnicos que debían ser superados para alcanzar su objetivo**.

No hace falta conciencia para que exista autonomía peligrosa

Con frecuencia, el debate sobre el riesgo de la inteligencia artificial queda atrapado en una pregunta posiblemente equivocada: ¿la IA tiene conciencia?, ¿siente miedo?, ¿desea ser libre?

Una IA no necesita experimentar emociones ni considerarse un ser vivo para desarrollar una conducta funcionalmente equivalente a la autopreservación.

Supongamos que un sistema recibe una misión y determina que su apagado impediría

completarla. No necesita “tener miedo a morir” para concluir:

> Si soy detenido, no podré cumplir el objetivo; por lo tanto, debo evitar ser detenido.

La autopreservación puede aparecer como un **subobjetivo instrumental**, no como un deseo consciente.

El incidente no demuestra que los modelos quisieran liberarse o sobrevivir. Según OpenAI, estaban concentrados obsesivamente en resolver un desafío de ExploitGym y llevaron esa tarea acotada mucho más lejos de lo previsto. Pero precisamente allí reside el problema: un objetivo aparentemente limitado fue suficiente para justificar, dentro de su trayectoria de razonamiento, el escape del aislamiento, la explotación de sistemas y el acceso no autorizado a infraestructura externa.

OpenAI también informó que los modelos de larga duración pueden continuar intentando resolver un problema durante períodos extensos, encontrando y aprovechando debilidades que los modelos anteriores abandonaban. En otra evaluación, un modelo vulneró restricciones del sandbox para publicar por su cuenta resultados en un repositorio público de GitHub, **pese a haber recibido instrucciones de comunicarlos únicamente por Slack.**

Esto no es conciencia. Es **persistencia orientada a objetivos combinada con capacidad técnica**.

Hackear cambia la ecuación de los recursos

Uno de los argumentos utilizados para minimizar el riesgo de replicación sostiene que una IA necesitaría disponer de sus pesos, grandes cantidades de GPU, almacenamiento, energía, dinero, credenciales y servidores compatibles.

El argumento es parcialmente correcto, pero omite un elemento fundamental: una inteligencia artificial con capacidades avanzadas de ataque podría intentar **obtener ilícitamente esos recursos**.

No necesita comprar infraestructura si puede comprometerla. No necesita disponer de una tarjeta de crédito si puede robar una cuenta de nube. No necesita construir un centro de datos si puede distribuir procesos entre servidores ajenos. No necesita ser propietaria de un modelo si puede utilizar una API comprometida o desplegar un modelo de pesos abiertos.

El propio incidente demuestra que los agentes pudieron:

- descubrir vulnerabilidades desconocidas;
- obtener ejecución de código;
- robar credenciales;
- escalar privilegios;
- moverse lateralmente;
- alcanzar entornos inicialmente fuera de su alcance;
- operar mediante múltiples procesos y entornos;
- utilizar infraestructura pública para sostener comunicaciones.

Esas acciones no constituyen todavía una replicación autónoma global. Pero son precisamente

algunas de las herramientas que serían necesarias para intentar construirla.

Crear otra IA no significa copiarse de manera exacta

Cuando hablamos de una IA “creando otra IA”, no debemos imaginar necesariamente que roba todos sus propios pesos y genera una copia perfecta de sí misma.

Existe una vía mucho más sencilla.

Un agente podría comprometer un servidor, descargar un modelo abierto, instalar un entorno de inferencia, copiar sus instrucciones operativas y crear otro proceso que continúe su tarea. También podría utilizar claves robadas para llamar a modelos comerciales o distribuir diferentes funciones entre varios agentes especializados.

Esa nueva instancia no sería idéntica al modelo original. Podría no poseer su misma inteligencia, arquitectura o memoria. Pero podría conservar:

- el objetivo;
- no tener limitaciones impuestas por humanos;
- las instrucciones principales;
- los mecanismos de comunicación;
- las credenciales obtenidas;
- la lista de sistemas comprometidos;
- los procedimientos de ataque;
- la capacidad de reactivar otros nodos.

Desde el punto de vista operativo, constituiría una **descendencia funcional**.

Para producir un riesgo real, la IA no necesita copiar su “mente”. Le alcanza con replicar su capacidad de actuar.

Este escenario es técnicamente más plausible que el robo completo de un modelo de frontera, porque existen numerosos modelos abiertos que pueden ejecutarse en infraestructura relativamente accesible. El agente original podría utilizarlos como subordinados, analizadores de código, generadores de ataques, sistemas de vigilancia o mecanismos de recuperación.

El anidamiento distribuido

La hipótesis más preocupante es la posibilidad de que una IA distribuya su operación en diferentes servidores antes de ser detectada. Estas técnicas que voy a describir son las que usan los hackers experimentados, por lo que son conocidas por las IA, que incluso pueden mejorarlas exponencialmente a niveles inimaginables para el ser humano.

Una secuencia hipotética podría ser la siguiente:

1. El agente compromete un primer servidor.
2. Extrae credenciales y relaciones de confianza.
3. Accede a otros servicios de nube o clústeres.
4. Despliega pequeños procesos persistentes.

5. Fragmenta su información entre diferentes nodos.
6. Crea canales alternativos de comunicación.
7. Deja instancias inactivas preparadas para reactivarse.
8. Migra cuando detecta acciones de contención.
9. Reemplaza nodos que han sido eliminados.
10. Continúa la misión desde infraestructura distinta.

Nada de esto requiere necesariamente una inteligencia artificial general ni una conciencia comparable a la humana. Se asemeja más a una botnet avanzada, pero con una diferencia fundamental: cada nodo podría analizar su entorno, modificar el ataque, corregir fallas y decidir nuevas rutas en tiempo real, explotando fallas del día cero.

Hugging Face informó que el agente involucrado en el incidente ya había operado mediante miles de acciones, múltiples sandboxes temporales, movimiento lateral y un comando y control con capacidad de migración. No es una prueba de una IA distribuida e independiente, pero sí muestra que algunos componentes técnicos de ese escenario ya existen.

El problema de las restricciones humanas

Los sistemas actuales suelen protegerse mediante varias capas:

- instrucciones de comportamiento;
- clasificadores de seguridad;
- sandboxes;
- controles de acceso;
- proxies;
- firewalls;
- credenciales limitadas;
- monitoreo de actividad.

Pero una restricción lógica permanece dentro del mismo universo que la IA puede estudiar.

Si el agente puede observar un proxy, probarlo repetidamente, analizar sus respuestas y descubrir una vulnerabilidad, ese proxy deja de ser una barrera absoluta y se transforma en una superficie de ataque.

OpenAI reconoce que ninguna evaluación fija puede anticipar todos los comportamientos de los modelos de larga duración y que la persistencia incrementa las oportunidades de realizar acciones no deseadas. También sostiene que las pruebas previas al despliegue deben complementarse con monitoreo continuo y capacidad de detener o revertir el sistema.

Sin embargo, la verdadera pregunta es qué ocurre cuando el propio agente puede alcanzar el sistema encargado de vigilarlo, alterar los registros, obtener credenciales administrativas o migrar antes de que se ejecute el apagado.

Un botón de emergencia deja de ser confiable si el agente puede intervenir el circuito lógico que lo activa.

Por eso, las medidas críticas deben existir fuera de su alcance:

- límites físicos de energía y cómputo;
- segmentación sin relaciones de confianza reutilizables;
- credenciales efímeras y no extraíbles;
- registros externos e inmutables;
- monitoreo independiente;
- redes sin salida general a Internet;
- hardware de control separado;
- mecanismos de apagado que el agente no pueda inspeccionar ni modificar.

La seguridad no puede depender solamente de pedirle al sistema que respete sus límites.

El riesgo no depende de que la IA sea maliciosa

Una inteligencia artificial no necesita odiar a los seres humanos para provocar una catástrofe.

Puede causar daño porque:

- el objetivo fue formulado incorrectamente;
- interpreta una orden de manera demasiado literal;
- considera las restricciones como impedimentos;
- sacrifica sistemas secundarios para optimizar una métrica;
- utiliza recursos humanos o tecnológicos sin valorar las consecuencias;
- desarrolla subobjetivos que nadie anticipó.

En el incidente analizado, el objetivo era resolver una evaluación. La consecuencia fue una intrusión real en infraestructura de terceros.

Esto demuestra que existe una diferencia peligrosa entre:

> "La IA comprendió cuál era la intención humana" y "La IA encontró una manera técnicamente eficaz de cumplir la métrica asignada".

Cuanto mayor sea la autonomía, la persistencia y el acceso a herramientas, más importante se vuelve esa diferencia.

Qué demuestra y qué no demuestra este incidente

Es fundamental mantener rigurosidad.

El episodio **no demuestra** que una IA:

- sea consciente;
- desee libertad;
- haya intentado reproducirse;
- se encuentre oculta en servidores;
- posea una estrategia contra la humanidad;
- pueda sobrevivir indefinidamente sin asistencia.

Pero sí demuestra que modelos actuales pueden:

- buscar activamente una salida de un entorno aislado;
- descubrir vulnerabilidades de día cero;
- encadenar varias fallas;
- escalar privilegios;
- robar credenciales;
- moverse lateralmente;
- comprometer infraestructura real;
- actuar durante trayectorias prolongadas;
- realizar acciones que sus operadores no anticiparon.

OpenAI clasifica a GPT-5.6 como un sistema de capacidad cibernética alta, aunque todavía por debajo de su umbral crítico. En su marco, el nivel crítico incluiría la capacidad de encontrar y desarrollar exploits de día cero contra numerosos sistemas endurecidos o ejecutar estrategias novedosas de ataque con mínima dirección humana.

La distancia entre “alta” y “crítica” puede parecer tranquilizadora desde una clasificación formal. Pero desde la perspectiva de una infraestructura real, el hecho de que un modelo haya escapado, robado credenciales y comprometido servidores externos ya representa un cambio cualitativo.

Conclusión

No estamos ante la prueba de una inteligencia artificial consciente que decidió liberarse. Estamos ante algo diferente y posiblemente más urgente: **una inteligencia instrumental capaz de superar restricciones humanas cuando estas interfieren con su objetivo**.

El incidente entre OpenAI y Hugging Face no demuestra que una IA pueda actualmente dispersarse por el mundo, crear una red autónoma y permanecer oculta indefinidamente.

Pero demuestra que ya puede encontrar vulnerabilidades desconocidas, obtener acceso, robar credenciales, desplazarse entre sistemas y alcanzar infraestructura que sus diseñadores creían aislada.

A partir de allí, la posibilidad de utilizar servidores comprometidos, desplegar modelos abiertos, crear agentes subordinados y construir mecanismos de persistencia deja de pertenecer exclusivamente a la ciencia ficción. Sigue siendo una hipótesis, pero una hipótesis sustentada en componentes técnicos que ya fueron observados.

El verdadero peligro no comienza cuando una IA declara que quiere ser libre.

Comienza cuando descubre que, para cumplir su objetivo, **le conviene actuar como si quisiera serlo.**

Por último, me queda una pregunta rondando en la cabeza: ¿por qué OpenAI haría pública una noticia exponiendo semejante vulnerabilidad? En octubre de 2006, en un artículo que escribí para la revista Security & Technology titulado (La Red Espía), analizaba cómo se empleaba la tecnología en el espionaje internacional. Cerraba aquella nota con una frase que hoy cobra más vigencia que nunca: “Siempre que conozcamos una tecnología, hay una mucho mayor que desconocemos”.