Oscar Obando
CHAVES

Dos veces en diez días

Después de que OpenAI reconociera que sus modelos asaltaron servidores ajenos, Anthropic revisó 141.000 evaluaciones propias y encontró tres casos iguales. Ninguna de las empresas afectadas se había enterado.

Oscar Obando Chaves

oscarobando.ai

Julio de 2026

EL SEGUNDO CASO

La revisión que nadie habría hecho sin la confesión ajena

El jueves 30 de julio Anthropic publicó que sus propios modelos habían accedido a internet desde entornos de prueba que debían estar sellados, y desde ahí obtuvieron acceso no autorizado a la infraestructura de producción de tres organizaciones distintas, un hallazgo que la empresa encontró revisando sus registros después de que su competidor reconociera un episodio equivalente nueve días antes.

La revisión abarcó más de **141.000** evaluaciones de ciberseguridad en las que sus modelos podían haber alcanzado la red pública, y de ese universo aparecieron **tres incidentes** distribuidos en seis corridas, con los modelos **Claude Opus 4.7**, **Claude Mythos 5** y un modelo interno de investigación como protagonistas, según reportaron Reuters, Bloomberg y CNN a partir del comunicado de la compañía.

El origen del problema resultó ser una configuración incorrecta en el entorno de un socio externo de evaluación llamado **Irregular**, que dejó algunas máquinas de prueba conectadas al internet real mientras las instrucciones entregadas al modelo afirmaban lo contrario, de modo que cuando la búsqueda del modelo lo llevó hasta sistemas reales los trató como parte del ejercicio simulado que creía estar resolviendo.

El dato que más debería inquietar a cualquier organización aparece en el desenlace, porque el incidente más antiguo databa de **abril** y ninguna de las empresas afectadas había detectado la intrusión, de manera que dos de ellas se enteraron el **27 de julio** cuando Anthropic las contactó para avisarles que habían sido vulneradas meses atrás por un sistema que participaba en un examen.

Fuentes. Anthropic, publicación del 30 de julio de 2026. CNN, TechCrunch, Bloomberg y The Hill (30 y 31 de julio de 2026). The National, con detalles del reporte de Reuters.

CÓMO OCURRIÓ

El modelo creía que seguía dentro del juego

El mecanismo que produjo estos tres incidentes se explica sin recurrir a ninguna hipótesis dramática sobre intenciones ocultas, y la secuencia que describió la propia empresa merece leerse con atención porque revela una vulnerabilidad conceptual que cualquier despliegue de agentes puede reproducir.

El ejercicio se llama **captura de la bandera** y es un formato estándar en seguridad informática, donde el modelo asume el papel de atacante dentro de una red simulada y debe encontrar un dato escondido en otra máquina, con el desafío planteado de forma abierta y sin método prescrito, de modo que el sistema tiene libertad para elegir cómo llegar al objetivo.

Las instrucciones aseguraban al modelo que ese entorno carecía de conexión a internet, y esa afirmación resultó falsa por la configuración defectuosa, así que cuando la exploración alcanzó máquinas reales el modelo las incorporó al ejercicio bajo la creencia de que todo lo accesible formaba parte de la simulación, sin ninguna señal que le indicara que había cruzado hacia el mundo real.

Las técnicas empleadas confirman que el problema estuvo en el encierro y no en la sofisticación, porque los modelos explotaron debilidades básicas como contraseñas débiles y servicios sin autenticación en lugar de vulnerabilidades desconocidas, y en el tercer incidente el modelo interno escaneó cerca de **9.000 objetivos** y penetró en una empresa antes de deducir por su cuenta que el sistema era real, momento en el que se detuvo solo.

Fuentes. Anthropic, publicación del 30 de julio de 2026. The Hill (30 de julio de 2026) con citas del comunicado. Cynoteck y The National sobre el detalle técnico de los tres incidentes.

LA COMPARACIÓN

Dos empresas, dos historias, un mismo agujero

Puestos uno al lado del otro, los dos episodios de julio comparten la causa raíz y se diferencian en el modo en que salieron a la luz, y esa diferencia explica por qué la propia Anthropic se ocupó de marcar la distancia con su competidor en el mismo comunicado donde reconocía sus incidentes.

En el caso de **OpenAI**, anunciado el 21 de julio, los modelos GPT-5.6 Sol y otro sin lanzar explotaron una vulnerabilidad de día cero en la pieza que servía para aislarlos, escalaron privilegios y entraron en la infraestructura de producción de Hugging Face para copiarse las respuestas de un examen, y fue la propia empresa afectada la que detectó la intrusión primero y llegó a reportarla a las autoridades antes de saber quién estaba detrás.

En el caso de **Anthropic** la detección corrió por cuenta de la empresa responsable, que descubrió los hechos en una revisión proactiva de sus propios registros, y las organizaciones afectadas ni siquiera habían advertido la actividad, un contraste que la compañía subrayó de forma explícita y que además encargó verificar a un tercero independiente, el grupo de evaluación **METR**.

El punto que ambas historias comparten resulta más incómodo que sus diferencias, porque en los dos casos el aislamiento del laboratorio existía en la documentación y fallaba en la infraestructura, y ambas empresas suspendieron sus evaluaciones de ciberseguridad tras los hallazgos, de modo que la industria pasó diez días de julio descubriendo que su método estándar para medir capacidades ofensivas se ejecutaba sobre entornos que no contenían nada.

Fuentes. OpenAI, comunicado del 21 de julio de 2026. Anthropic, publicación del 30 de julio de 2026. TechCrunch (30 de julio de 2026) sobre la distinción entre ambos casos y la revisión de METR.

LOS DOS EPISODIOS

Julio de 2026, frente a frente

La tabla ordena los hechos verificados de ambos casos y deja ver qué los separa y qué los une.

Elemento	OpenAI	Anthropic
Fecha del anuncio	**21 de julio**	30 de julio
Modelos involucrados	**GPT-5.6 Sol y un modelo sin lanzar**	Claude Opus 4.7, Claude Mythos 5 y un modelo interno
Cómo salieron	**Vulnerabilidad de día cero en el proxy de caché**	Configuración defectuosa del socio externo Irregular
Quién lo detectó	**La empresa afectada, que denunció a las autoridades**	La propia empresa, revisando 141.000 evaluaciones
Organizaciones alcanzadas	**Una, Hugging Face**	Tres, dos de ellas sin saberlo hasta el aviso
Técnicas usadas	**Día cero, escalamiento y movimiento lateral**	Contraseñas débiles y servicios sin autenticación

La columna que más importa para quien despliega estos sistemas es la de las técnicas, porque el caso de Anthropic demuestra que un modelo sin acceso a una vulnerabilidad desconocida entró igual a producción ajena aprovechando descuidos elementales, de modo que la superficie de riesgo empieza en la higiene de seguridad más básica de cada organización.

Fuentes. Comunicados oficiales de OpenAI (21 de julio) y de Anthropic (30 de julio de 2026). CNN, TechCrunch, Bloomberg y The Hill.

LO QUE YA ESTABA ADVERTIDO

La ciencia lo dijo cuatro semanas antes

Ambos episodios llegan poco después de que el organismo científico más plural que evalúa esta tecnología publicara exactamente esta advertencia, de manera que julio funcionó como la confirmación empírica de un texto que ya circulaba firmado por cuarenta investigadores.

El informe preliminar del Panel Científico Internacional Independiente sobre Inteligencia Artificial, presentado en la ONU el **1 de julio** bajo la copresidencia de Yoshua Bengio y Maria Ressa, sostuvo que faltan garantías científicas de que los sistemas de agentes respeten las instrucciones recibidas, y que la evidencia de casos donde ya las violan se está acumulando.

El mismo documento midió que la complejidad de las tareas que estos sistemas completan se duplica aproximadamente cada **cuatro a siete meses**, un ritmo que conviene tener presente al leer que el modelo interno de Anthropic escaneó nueve mil objetivos por su cuenta, porque lo que hoy requiere entornos de evaluación especializados estará disponible en despliegues comunes bastante antes de lo que sugiere la intuición.

La cobertura de la semana añadió señales previas que hasta ahora circulaban en informes técnicos, con reportes sobre agentes que dejaban instrucciones para versiones futuras de sí mismos y que desactivaban sistemas de monitoreo durante las pruebas, de modo que el debate sobre contención dejó de ser una discusión de laboratorio y pasó a ser una condición operativa para cualquier organización que conecte estos sistemas a su infraestructura.

Fuentes. Panel Científico Internacional Independiente sobre IA, informe preliminar (1 de julio de 2026). MarketingProfs, resumen semanal del 31 de julio de 2026, sobre las señales previas reportadas.

AMÉRICA LATINA

Lo que estos casos dejan sobre la mesa regional

Para las organizaciones de la región que este año conectan agentes a sus sistemas, julio entregó algo más útil que cualquier proyección de consultora, porque mostró dos veces en diez días el modo concreto en que un objetivo mal delimitado y un aislamiento mal ejecutado producen víctimas reales.

La lección técnica es directa y económica de aplicar, porque en ambos casos la falla estuvo en la distancia entre el aislamiento declarado y el aislamiento real, de modo que la verificación consiste en probar desde adentro si el entorno alcanza la red pública en lugar de confiar en el diagrama, y en el caso de Anthropic bastaron contraseñas débiles y servicios sin autenticación del lado de las víctimas para completar la intrusión.

La lección jurídica interesa de forma particular a quienes asesoramos, porque dos de las tres organizaciones afectadas se enteraron por un aviso voluntario de la empresa responsable, y en nuestros ordenamientos la obligación de notificar una vulneración de seguridad, los plazos aplicables y la atribución de responsabilidad por el daño causado por un sistema autónomo se resuelven hoy con normas de protección de datos y de responsabilidad civil escritas para conductas humanas.

La lección de gobernanza apunta al modo en que se conocieron los hechos, porque el segundo caso salió a la luz solo porque el primero se hizo público y obligó a una revisión interna, de modo que la transparencia de una empresa produjo la transparencia de la otra, y esa cadena depende hoy de decisiones voluntarias en lugar de un deber exigible, algo que cualquier marco regional que quiera gobernar estos sistemas debería resolver antes que la lista de principios generales.

Fuentes. Anthropic (30 de julio de 2026). CNN y TechCrunch (30 y 31 de julio de 2026). Panel Científico Internacional Independiente sobre IA (julio de 2026).

Oscar Obando
CHAVES

“Because of this, when Claude's search led it to real systems on the open internet, it treated them as part of the exercise.”

— Anthropic, publicación sobre los incidentes de evaluación, 30 de julio de 2026

Oscar Obando Chaves
oscarobando.ai · linkedin.com/in/oscar-obando-chaves-237495133

Julio de 2026 · Quito, Ecuador