PARTE 6

Los agentes de IA y la nueva carrera por la remediación

Reducir las rutas explotables antes de que los atacantes las encuentren.

Los anuncios recientes sobre habilidades avanzadas de IA, como Claude Mythos Preview de Anthropic, GPT 5.5 de OpenAI y MDASH de Microsoft, señalan un cambio en la seguridad ofensiva que resulta cada vez más difícil de ignorar. Los agentes de IA ahora pueden encontrar una vulnerabilidad, validarla y encadenarla con otras exposiciones para alcanzar objetivos de mayor valor, con una intervención humana mínima.

A medida que el acceso a estas capacidades se amplía, esta mayor autonomía puede ayudar a los atacantes a pasar del descubrimiento a la explotación más rápidamente de lo que los controles de seguridad por capas pueden detectar y responder. Este cambio ha generado reacciones conocidas: desde quienes esperan los peores escenarios hasta quienes descartan estas afirmaciones con escepticismo. Sin embargo, para quienes adoptan una visión pragmática, ya existe evidencia suficiente para actuar. Independientemente de la perspectiva, la tarea de los CISO sigue siendo la misma: encontrar, validar y remediar las exposiciones con mayor probabilidad de generar un impacto significativo.

Esto apunta a una respuesta práctica: utilizar capacidades de IA similares en condiciones controladas para comprender qué es probable que encuentren los adversarios. Con supervisión liderada por especialistas, controles claros y un uso estrictamente delimitado, los equipos de seguridad pueden enfocar sus esfuerzos de remediación en los activos de mayor riesgo.

Cuando los ataques van más allá de vulnerabilidades aisladas

¿Qué ha cambiado?

Muchos flujos de trabajo ofensivos impulsados por IA se detenían anteriormente después de validar una sola vulnerabilidad. Lo que está cambiando de manera más visible es el encadenamiento y la velocidad que este aporta. Ahora es posible incorporar más decisiones entre cada paso del proceso, de manera que el resultado de una explotación pueda informar la siguiente. En algunos casos, la secuencia puede extenderse a cadenas cortas de dos o tres vulnerabilidades, reflejando más fielmente cómo ocurren las intrusiones reales.

Dos mejoras en los modelos de frontera facilitan este encadenamiento: ventanas de contexto más amplias y mayor capacidad para encontrar información relevante dentro de grandes bases de código. Juntas, estas mejoras permiten analizar grandes repositorios de código en contexto y rastrear flujos de datos de extremo a extremo. Esto puede revelar vulnerabilidades que no son evidentes en fragmentos aislados, por ejemplo, cuando los datos se modifican en una parte de una aplicación y posteriormente se consideran confiables en otro lugar. Esta visibilidad ampliada aumenta el número de puntos de partida plausibles para el encadenamiento de vulnerabilidades.

¿Por qué es importante?

Históricamente, las organizaciones se han beneficiado del ritmo más lento del análisis de código y del encadenamiento automatizado de vulnerabilidades. A medida que más etapas de esta secuencia se automatizan, los atacantes pueden avanzar más rápido, ese margen de tiempo se reduce y el período disponible para la remediación deja de estar asegurado.

Esto cambia la forma de priorizar. Los líderes de seguridad deberían concentrarse menos en el volumen de hallazgos y más en utilizar modelos avanzados para comprender qué exposiciones podría encadenar y explotar un atacante. El desafío inmediato es acelerar la validación y la remediación para mantener el ritmo.

A medida que aumenta el encadenamiento: más hallazgos que verificar y menos tiempo para corregir

A medida que la capacidad de encadenar vulnerabilidades escala, el desafío de la clasificación y la remediación se convierte en un problema tanto de velocidad como de volumen. Los métodos automatizados de evaluación pueden generar hallazgos plausibles a partir de pruebas internas, divulgaciones públicas y programas de recompensas por vulnerabilidades. Sin embargo, cada posible problema aún requiere validación. Ese trabajo compite directamente con la capacidad necesaria para corregir vulnerabilidades confirmadas.

Esta dinámica ya es visible en ecosistemas de código abierto. Los mantenedores informan que están inundados de solicitudes de combinación de código y reportes de errores generados por IA de escaso valor, los cuales deben revisarse, clasificarse y, con frecuencia, rechazarse, ralentizando la corrección de vulnerabilidades críticas. El mismo patrón puede ocurrir dentro de las empresas. Los resultados de evaluaciones extensas pueden revelar muchas exposiciones potenciales, pero cada una debe validarse en contexto.

Cuando los hallazgos tienen poco valor o no son reproducibles, la verificación manual se transforma en un costo hundido que compite con la remediación y desvía tiempo de los esfuerzos para reducir riesgos reales.

Esto representa tanto un desafío tecnológico como operativo. Requiere una disciplina más sólida en la recepción, clasificación y validación de hallazgos para que los equipos mantengan el foco en los problemas con mayor probabilidad de generar impactos importantes.

Utiliza Confianza en la IA para identificar lo que encontrarán los adversarios

Si los atacantes pueden utilizar capacidades ofensivas avanzadas para identificar y conectar exposiciones, las organizaciones deberían utilizar esas mismas capacidades en condiciones controladas para comprender qué encontrarán y explotarán los adversarios. La diferencia está en el control. PwC denomina este enfoque Confianza en la IA (Trust AI).

A medida que los flujos de trabajo de los adversarios se vuelven más rápidos y más capaces de conectar diferentes etapas del ataque, las organizaciones enfrentan una elección práctica: tratar estas capacidades únicamente como una amenaza externa o utilizar capacidades similares internamente para mejorar la identificación, verificación y priorización de exposiciones. Este enfoque puede ampliar las capacidades de seguridad ofensiva sin reemplazar pruebas de penetración o ejercicios de *red team* de alto valor, ayudando a identificar exposiciones conectadas antes y transformando esos conocimientos en acciones de remediación.

Los líderes tienen razones válidas para ser cautelosos respecto al uso de herramientas autónomas en entornos de producción sin tener confianza en lo que hará el sistema, cómo se hará cumplir el alcance o si la actividad puede detenerse inmediatamente. El valor surge cuando estos métodos se aplican siguiendo los principios de Trust AI de forma controlada, por equipos capaces de gobernar las herramientas, interpretar los resultados y comprender el entorno que se está evaluando.

Un modelo controlado suele implicar el establecimiento de reglas de participación claras, permisos estrictamente limitados, y un mecanismo de supervisión humana continua que permita pausar o detener la actividad si el comportamiento se desvía de su propósito previsto. También prioriza la repetibilidad: primero probar en condiciones restringidas y luego ampliar el alcance únicamente cuando los resultados puedan ser confiables y aplicables.

Una interrupción ampliamente reportada en un importante proveedor de servicios en la nube ilustra por qué una gobernanza adecuada es tan importante.

Cuando los sistemas autónomos operan con privilegios amplios debido a reglas de acceso mal configuradas, incluso las acciones rutinarias pueden generar consecuencias significativas.

La supervisión humana es importante por una razón más simple: estos sistemas pueden equivocarse con total convicción. Sin personas capaces de validar los resultados, cuestionar las suposiciones y traducir los hallazgos técnicos en impacto de negocio, las organizaciones corren el riesgo de desarrollar una falsa confianza y desperdiciar esfuerzos corrigiendo los problemas equivocados. El objetivo es reducir las rutas viables hacia el impacto, no generar más hallazgos.

Qué hacer a continuación

Pon a prueba las suposiciones dependientes del tiempo

Identifica los controles y procesos de los que dependes y que asumen largos períodos de detección o una lenta progresión por parte de los atacantes. Determina dónde es más probable que se formen exposiciones conectadas dentro de tu entorno: aplicaciones críticas, rutas con privilegios elevados y servicios expuestos externamente. Luego confirma responsables claros y rutas de remediación definidas.

Refuerza la disciplina de validación y clasificación

Eleva los estándares de calidad para la recepción de hallazgos de modo que los reportes de bajo esfuerzo no se conviertan en investigaciones de alto costo. Separa los hallazgos plausibles de los confirmados mediante un flujo de trabajo que escale únicamente aquello que pueda demostrarse en tu entorno. Consolida señales repetidas provenientes de escáneres, programas de recompensas y canales de divulgación en un único elemento listo para ingeniería. Prioriza las exposiciones conectadas concentrándote en los hallazgos que generen una ruta hacia un activo material, y no únicamente en aquellos con la mayor severidad nominal. Realiza seguimiento al tiempo de validación y al tiempo dedicado por especialistas a cada problema validado para que el costo sea visible y gestionable.

Aplica una perspectiva de costo e impacto a las decisiones sobre herramientas

Utiliza modelos y enfoques capaces de descubrir aquello que las herramientas estándar o las pruebas de penetración tradicionales podrían pasar por alto, sin permitir que el costo de validación supere la capacidad de remediación. Evalúa qué es lo adecuado para tu organización en función del costo de uso, los riesgos que intentas reducir y el nivel de capacidad que realmente necesitas.

Implementa pruebas ofensivas controladas

Comienza con un alcance limitado y reglas de participación claras. Aplica controles de supervisión humana continua. Realiza pruebas repetibles antes de ampliar el alcance. Enfoca los resultados en acciones de remediación respaldadas por evidencias que reduzcan las exposiciones explotables, en lugar de maximizar la cantidad de hallazgos. Después, incorpora los patrones recurrentes a tus estándares de ingeniería y a tus mecanismos de gobernanza.

Los tiempos de los ataques se están reduciendo. Los defensores que validen eficazmente, utilicen métodos avanzados de seguridad ofensiva con supervisión clara y cierren las rutas con mayor probabilidad de ser explotadas estarán mejor posicionadas para mantener el ritmo.