INFORME

Seguridad de la IA 2026

Observaciones realizadas en una base de clientes compuesta por más de 625 000 organizaciones de todo el mundo

SOPHOS

SOPHOS | X-OPS

Carta de John Peterson

En mayo de 2026, Sophos descubrió a un atacante que llevaba a cabo lo que equivalía a una operación de desarrollo de software dentro de la red de un cliente. Alrededor de una docena de agentes de IA, coordinados mediante un asistente de programación comercial, escribían y probaban malware contra la protección para endpoints de Sophos, CrowdStrike y Windows Defender, cada uno en su propia máquina virtual. La operación produjo casi 80 módulos, más de 70 técnicas de evasión y cada resultado se incorporaba al control de versiones y se mejoraba en la siguiente iteración. Posteriormente, el mismo operador utilizó los resultados para desplegar ransomware y robar datos.

Aunque las técnicas empleadas no eran nuevas, los agentes convirtieron semanas de iteración manual en días de iteración automatizada, y este cambio es el tema central del presente informe. La IA está cambiando la velocidad y la forma de las operaciones de seguridad más de lo que está cambiando los fundamentos de las intrusiones. Los atacantes siguen necesitando acceso inicial, siguen moviéndose lateralmente y siguen exfiltrando datos a través de canales observables. Lo que ha cambiado es el tiempo.

En este informe, Sophos demostrará esta afirmación con datos.

Contamos con analistas de los servicios de detección y respuesta gestionadas (MDR) y de respuesta a incidentes (IR) que investigan miles de incidentes al año. Contamos con la Counter Threat Unit (CTU), que supervisa foros clandestinos en varios idiomas. Contamos con investigadores de SophosLabs que analizan muestras de malware a gran escala. Contamos con un equipo de Sophos AI que desarrolla técnicas innovadoras orientadas a la defensa. Y contamos con datos de telemetría de endpoints, firewalls, correo electrónico, nube, red e identidad procedentes de una base global de más de 625 000 clientes.

Esta escala, alcance y experiencia nos proporcionan una perspectiva excepcional sobre la intersección en rápida evolución entre la IA y la ciberseguridad. Tenemos la intención de utilizar esta perspectiva para ayudar al sector a adelantarse a los acontecimientos y garantizar que nuestros clientes sigan protegidos.

J. P. Peterson, CTO, Sophos

Resumen ejecutivo y principales conclusiones

Este informe se basa en los casos de Sophos MDR e IR, los análisis de SophosLabs, la información de Sophos CTU, así como en las observaciones relacionadas con los endpoints y las redes de una base de clientes de más de 625 000 organizaciones de todo el mundo. Un mismo patrón aparece en todos ellos: La IA permite tanto a los atacantes como a los responsables de la seguridad llevar a cabo operaciones habituales con mucha mayor rapidez. La evidencia de tipos de ataque totalmente nuevos sigue siendo escasa, aunque no nula, y es muy probable que vaya a más conforme evolucionen sus capacidades.

2026 marca un punto de inflexión. Los modelos de frontera y los modelos de pesos abiertos ya resultan suficientemente útiles para delegar tareas de ingeniería, mientras que la reducción de los costes de inferencia modifica la viabilidad económica para ambas partes. La decisión práctica ya no consiste en determinar si se utilizará la IA, sino dónde se justifican los costosos sistemas de vanguardia, dónde bastan modelos de pesos abiertos más económicos y cómo distribuir el trabajo sin aumentar el riesgo.

Las siguientes conclusiones constituyen la base del informe:

1. **IA está reduciendo la duración de los ataques, no inventando nuevos tipos de ataque.** El ciberdelincuente realiza un seguimiento como STAC6994 utilizó agentes de IA para perfeccionar técnicas destinadas a eludir la EDR a un ritmo que habría requerido semanas para un desarrollador humano. Posteriormente, los analistas de Sophos CTU confirmaron que el operador responsable de STAC6994 llevó a cabo operaciones de despliegue de ransomware y robo de datos.
2. **La capa de credenciales e identidad relacionada con los servicios de IA para empresas se ha convertido ahora en un objetivo de recopilación de datos.** El incidente de Salesloft/Drift demostró cómo los tokens OAuth de los chatbots pueden conducir directamente a la vulneración de entornos de Salesforce. Por otra parte, los analistas de Sophos CTU observaron afirmaciones de ciberdelincuentes en foros clandestinos según las cuales las instrucciones de IA y los datos generados podrían capturarse de forma colateral durante los ciberataques.
3. **La economía sumergida está incorporando la IA como infraestructura.** Sophos documentó la venta por parte de ciberdelicuentes de claves de API obtenidas a través de intermediarios para Claude, ChatGPT y Grok. En los foros han aparecido canales dedicados a la ingeniería de instrucciones de IA, las técnicas de jailbreaking y los flujos de trabajo de desarrollo de malware. Algunas empresas están contratando ingenieros de instrucciones de IA. Otros siguen mostrándose abiertamente escépticos respecto a que la IA vaya a cambiar sus operaciones. La curva de adopción se asemeja menos a una transformación repentina y más a la incorporación gradual de cualquier herramienta nueva a los flujos de trabajo delictivos existentes
4. **Los ataques a la cadena de suministro se dirigen específicamente contra la infraestructura de desarrollo de IA.** La vulneración de la extensión Nx Console para VS Code distribuyó malware para el robo de credenciales. Los analistas de SophosLabs investigaron páginas falsas de descarga de Claude que distribuían un nuevo troyano de acceso remoto (RAT), así como una campaña ClickFix que utilizaba URL legítimas de chats compartidos de ChatGPT para distribuir el infostealer MacSync.

5. **La ingeniería social asistida por IA ha pasado de la fase experimental a la operativa.** Sophos CTU documentó anuncios en foros clandestinos de bots de voz, perfiles generados por IA para fraudes románticos y servicios de imágenes de perfil sintéticas. La aportación de la IA consiste en aumentar la escala y la calidad lingüística en distintos idiomas, aplicadas a las mismas técnicas de engaño que ya funcionaban antes de la IA generativa.

6. **Los plazos de explotación se están reduciendo más deprisa que los plazos de aplicación de parches.**
La Directiva Operativa Vinculante 26-04 de CISA, publicada el 10 de junio de 2026, redujo a tres días el plazo obligatorio para aplicar parches a las vulnerabilidades de mayor riesgo, mencionando expresamente la IA como uno de los factores que están reduciendo el tiempo entre la divulgación y la explotación. La primera prueba de la directiva en situaciones reales tuvo lugar casi de inmediato: CVE-2026-10520 (Ivanti Sentry) fue explotada en las 24 horas siguientes a la publicación de la prueba de concepto y se añadió al catálogo KEV el 12 de junio. CVE-2026-42208 (LiteLLM) fue explotada en un plazo de 36 horas, y los atacantes se dirigieron contra bases de datos que contenían claves de proveedores de LLM externos.

7. **Demostrar el uso de la IA en un ataque sigue resultando difícil desde el punto de vista operativo.** El caso de STAC6994 fue inusual porque el marco de desarrollo quedó en un dispositivo que Sophos pudo examinar. En muchos incidentes, el uso de la IA no es visible en los datos de telemetría. Es probable que la prevalencia real de los ataques asistidos por IA sea superior a lo que cualquier proveedor puede demostrar directamente.

8. **Los modelos de razonamiento de IA aceleran la investigación.** Los modelos de razonamiento de IA reducen a minutos las horas de investigación de los analistas cuando se trata de patrones de ataque conocidos. No sustituyen a los sistemas de detección de comportamientos que localizan una única amenaza oculta entre miles de millones de eventos. Las limitaciones estadísticas derivadas del desequilibrio extremo entre clases, la necesidad de contar con referencias específicas para cada entorno y la rapidez con la que se adaptan los adversarios requieren una ingeniería especializada.

Parte 1:
La IA en la cadena de ataque

Taxonomía de amenazas basadas en IA de Sophos X-Ops

Sophos X-Ops divide las amenazas basadas en IA en dos categorías principales: el uso malicioso de la IA, es decir, las amenazas frente a las que protegemos, y los ataques dirigidos contra la IA (la IA que protegemos). La primera incluye los artefactos generados por IA, las capacidades en tiempo de ejecución ampliadas mediante IA y las intrusiones coordinadas por IA. La segunda incluye las vulneraciones iniciadas por agentes, la suplantación de software de IA, los LLM contaminados y los ataques dirigidos contra los propios modelos. Ambas categorías requieren respuestas diferentes: el uso malicioso supone un problema de productividad y escala, mientras que los ataques dirigidos contra la IA plantean un problema de confianza, cadena de suministro y gobernanza. La taxonomía complementa marcos existentes como MITRE ATLAS, la taxonomía de Microsoft sobre modos de fallo de los agentes, el AI Risk Repository del MIT y NIST AI 100-2.

Figura 1: Descripción general de la taxonomía de las amenazas basadas en IA

Un gradiente de autonomía

En el nivel más básico, los ataques generados por IA consisten en que una persona utiliza IA generativa para producir un artefacto y lo despliega manualmente. Un ciberdelicuente que atacaba a organismos gubernamentales mexicanos utilizó Claude Code y GPT para generar scripts. Las conversaciones filtradas del grupo de ransomware The Gentlemen confirmaron usos similares. Además, en un caso reciente de ransomware DragonForce investigado por el equipo de Sophos IR, el ciberdelicuente elaboró lo que consideramos, con un alto grado de certeza, un informe generado por IA durante las negociaciones con la organización atacada. Este contenía un análisis detallado de los datos exfiltrados, incluida información de identificación personal (PII) y riesgos legales. El análisis generado por IA se utilizó como parte de un intento de ejercer presión y persuadir a la organización para que pagara el rescate.

Más adelante en el gradiente, los ataques generados por IA utilizan un modelo para mejorar una capacidad durante la ejecución. LameHug, un malware basado en Python que CERT-UA atribuye con un grado de confianza moderado a IRON TWILIGHT (APT28), consultaba un modelo de pesos abiertos alojado para generar dinámicamente comandos de reconocimiento de Windows adaptados a cada entorno. Esto reducía considerablemente la utilidad del análisis estático y convertía el tráfico saliente hacia endpoints de API de IA/ML en una de las pocas superficies de detección fiables. La suplantación ampliada mediante IA constituye otro vector activo: las herramientas de clonación de voz e intercambio de rostros ya se utilizan para cometer fraudes en tiempo real, suplantar a directivos y eludir los procesos KYC.

En el extremo más avanzado, la divulgación de Anthropic de noviembre de 2025 documentó una campaña patrocinada por el Estado chino que utilizó Claude Code para intentar vulnerar aproximadamente treinta objetivos. Estos ataques siguen siendo poco frecuentes, pero reducen el tiempo entre el reconocimiento y la acción de formas que ponen en entredicho los supuestos tradicionales de la defensa.

Estudio de caso: *STAC6994*

Los analistas de Sophos observaron a un ciberdelicuente utilizando IA para probar tácticas de evasión de la EDR en un marco de "Red Team" posterior a la explotación y descubrieron un dispositivo no autorizado que generaba cargas útiles de forma continua en el entorno de un cliente. El atacante había aprovisionado máquinas virtuales desde Ludus y utilizaba el IDE Cursor con agentes de IA para desarrollar y probar herramientas posteriores a la explotación.

El marco ejecutaba pruebas en paralelo en tres entornos: una máquina virtual con protección para endpoints de Sophos, otra con CrowdStrike y otra sin ninguna solución EDR como entorno de control. Una cuarta máquina virtual funcionaba como servidor de comando y control (C2) Sliver. Aproximadamente 12 agentes de IA desempeñaban funciones definidas.

Un agente que ejecutaba Claude Opus 4.5 se encargaba de las operaciones principales y del establecimiento de reglas. Otros gestionaban el refuerzo de la seguridad operativa (OPSEC), la documentación, las pruebas de resistencia de proxies, el despliegue de máquinas virtuales y las pruebas de las herramientas frente a los agentes EDR. Los commits se enviaban a Git mediante el Protocolo de Contexto de Modelo (MCP).

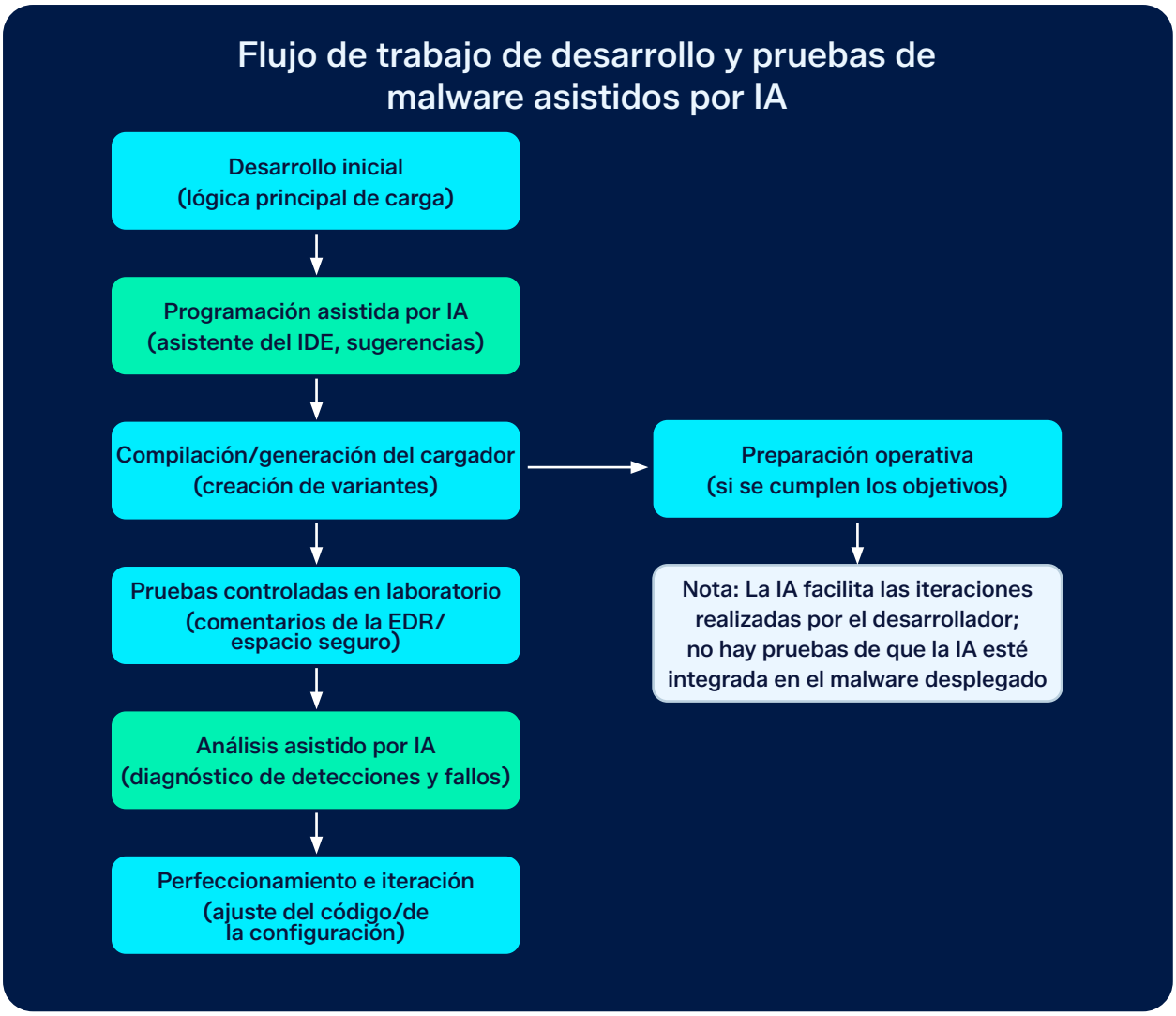


Figura 2: Diagrama que muestra la función de la IA en el flujo de trabajo de desarrollo de malware

La estrategia era sistemática. Según los artefactos recuperados, los agentes leían artículos de investigación extraídos del blog de SpecterOps, identificaban técnicas ofensivas, las asociaban con el marco MITRE ATT&CK, determinaban los pasos y las herramientas necesarios para reproducir cada técnica, preparaban el entorno de laboratorio, ejecutaban las pruebas y comunicaban los resultados.

El elemento central era un generador modular de cargas útiles basado en Python que envolvía las cargas útiles sin procesar en capas de cifrado y técnicas de evasión para producir ejecutables personalizados en Rust y Go. Se desarrollaron casi 80 módulos diferentes para probar más de 70 técnicas distintas. La ruta real para eludir la EDR consistía en un ciclo estructurado de pruebas de ingeniería con revisión e iteración humanas. La IA coordinaba el flujo de trabajo y facilitaba la experimentación; las pruebas no demostraron que la IA estuviera integrada en el malware desplegado.

Tras varias iteraciones, la documentación de los agentes afirmaba haber logrado un éxito prácticamente absoluto frente a los agentes EDR, aunque los analistas de Sophos señalaron que las pruebas no respaldaban plenamente esa conclusión. La señal operativa crucial fue el ritmo: el ciclo completo redujo de semanas a días el tiempo necesario.

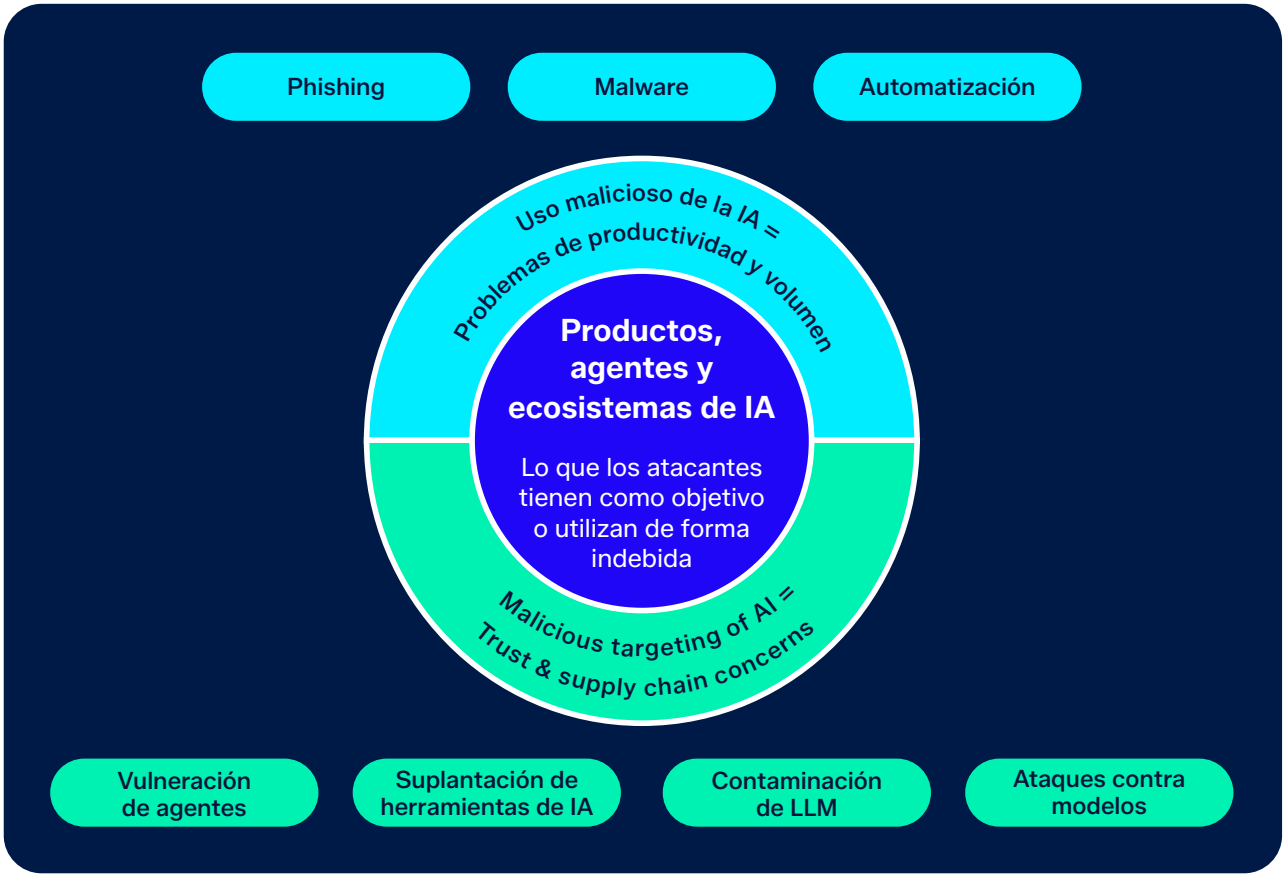
La CTU confirmó posteriormente que el operador de STAC6994 participó en operaciones de despliegue de ransomware y robo de datos; el marco era una herramienta de producción utilizada en intrusiones reales.

Ataques maliciosos dirigidos contra la IA

La segunda parte de nuestra taxonomía abarca los ataques en los que los productos, agentes y ecosistemas de IA constituyen el objetivo o la superficie de ataque, en lugar de ser la herramienta, y los analizaremos con más detalle en la parte 2. La vulneración iniciada por agentes es uno de los subtipos que suscitan una preocupación más inmediata: por ejemplo, un agente de programación que descarga un paquete NPM vulnerado o interactúa con un servidor MCP contaminado mientras realiza su trabajo. En este caso, la preocupación radica en la reducción del intervalo entre la publicación y la ejecución, sin intervención humana para revisar el registro de cambios.

Nuestra taxonomía también incluye la suplantación de software de IA, en la que los ciberdelicuentes se aprovechan de los usuarios que buscan herramientas de IA legítimas; la contaminación de los LLM, mediante la configuración de un modelo personalizado o la introducción intencionada de contenido malicioso o engañoso para modificar el comportamiento de un modelo; y diversos ataques contra los propios modelos, como la extracción de modelos, la inversión de datos de entrenamiento, los ejemplos de adversarios y la inferencia de pertenencia.

Tratar por separado las dos categorías principales ofrece una visión más clara de la amenaza. El uso malicioso de la IA es fundamentalmente un problema de productividad y volumen; las soluciones de detección actuales pueden identificar gran parte de esta actividad, pero el problema reside en la escala y la velocidad de iteración. Los ataques maliciosos dirigidos contra la IA suponen un problema de confianza y de cadena de suministro que resulta más adecuado abordar mediante políticas, marcos de diseño seguro e iniciativas similares.



Adopción y escepticismo en los foros clandestinos

Sophos lleva siguiendo las actitudes hacia la IA generativa en foros delictivos desde 2023, con un seguimiento en 2025 y una supervisión continua. El panorama ha evolucionado considerablemente. Los ciberdelicuentes compran claves de API a través de intermediarios para ChatGPT, Claude y Grok, y en los foros han surgido canales dedicados a la IA donde distintos perfiles comparten plantillas de instrucciones y técnicas de jailbreaking. Los investigadores de Sophos CTU señalaron que un conocido reclutador de los foros clandestinos, al que anteriormente habían observado contratando desarrolladores de blockchain y personal de centros de llamadas para ataques de phishing, intentaba contratar específicamente a un ingeniero de instrucciones de IA, externalizando los conocimientos especializados en IA del mismo modo que los ciberdelicuentes externalizan el desarrollo de malware y el acceso inicial.

La CTU también observó anuncios de bots de voz con IA diseñados para el vishing y el fraude telefónico, con reseñas positivas de usuarios que indican que los bots pueden imitar de forma convincente el tono, la cadencia y los patrones conversacionales. El perfil "HackingRealm" anunciaba un servicio de "modelos de OnlyFans creados con IA" para fraudes románticos e ingeniería social, que generaba imágenes sintéticas para perfiles y contenido conversacional con el fin de crear perfiles escalables en plataformas de citas y redes sociales, además de un sitio web de apoyo y una página de comentarios con reseñas positivas.

Están apareciendo herramientas comercializadas como "dirigidas por IA": Leak Bazaar, para clasificar mediante Machine Learning los datos robados, y una versión actualizada de Cobalt Strike con integración mediante API REST y servidor MCP. Los adversarios están incorporando integraciones con LLM en flujos de trabajo conocidos La referencia a Claude en el anuncio de Cobalt Strike funciona en parte como señal de modernidad, pero también refleja que la integración con LLM se está convirtiendo en un elemento diferenciador en los mercados delictivos, aunque aporte principalmente mayor comodidad en lugar de técnicas más sofisticadas.

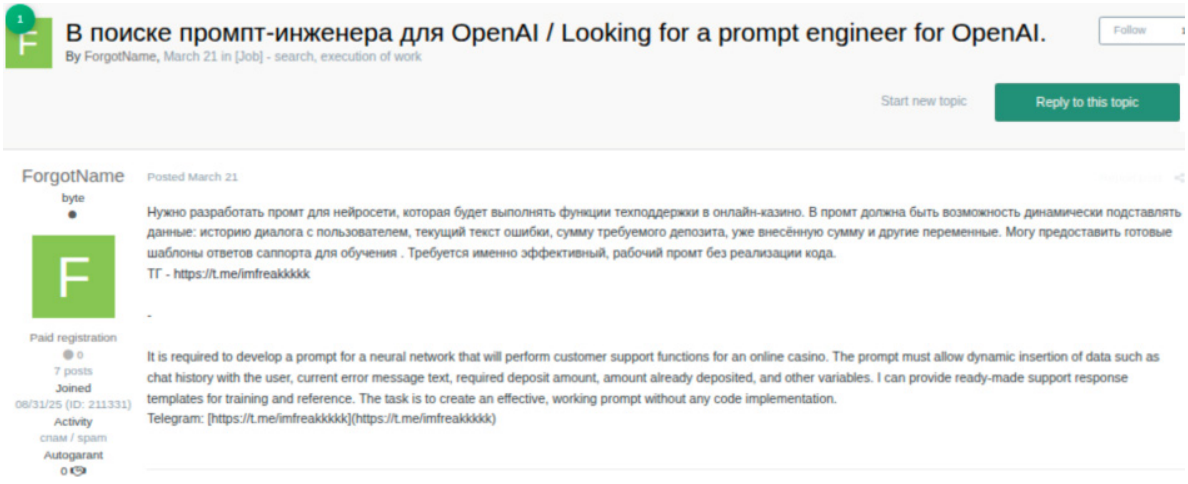
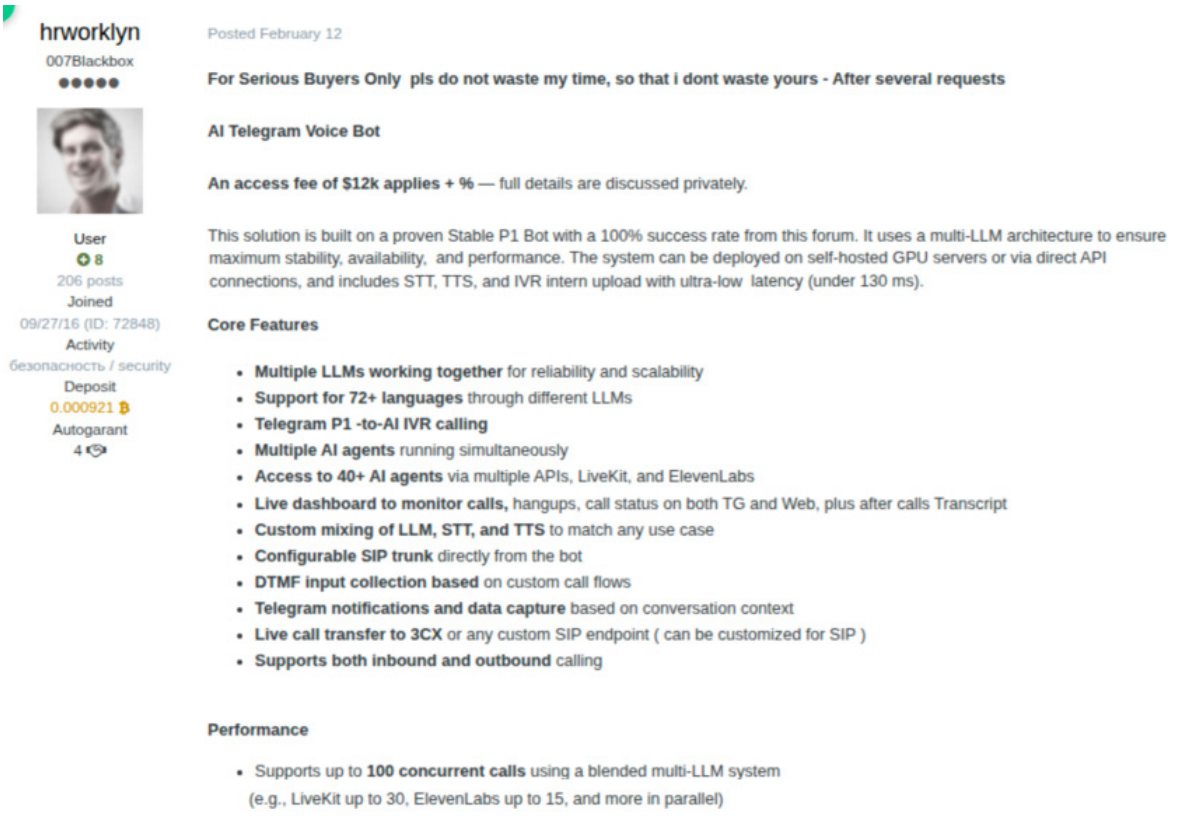


Figura 3: Anuncio de contratación de un ingeniero de instrucciones de OpenAI



hrworklyn
007Blackbox
●●●●●

User
8
206 posts
Joined
09/27/16 (ID: 72848)
Activity
безопасность / security
Deposit
0.000921 ₿
Autogarant
4

Posted February 12

For Serious Buyers Only pls do not waste my time, so that i dont waste yours - After several requests

AI Telegram Voice Bot

An access fee of $12k applies + % — full details are discussed privately.

This solution is built on a proven Stable P1 Bot with a 100% success rate from this forum. It uses a multi-LLM architecture to ensure maximum stability, availability, and performance. The system can be deployed on self-hosted GPU servers or via direct API connections, and includes STT, TTS, and IVR intern upload with ultra-low latency (under 130 ms).

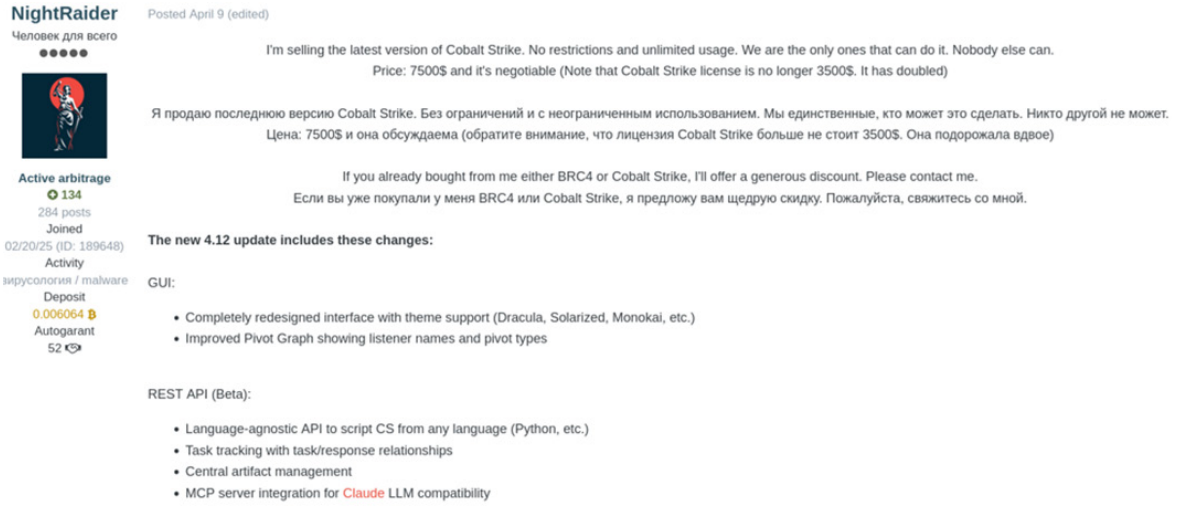
Core Features

- **Multiple LLMs working together** for reliability and scalability
- **Support for 72+ languages** through different LLMs
- **Telegram P1 -to-AI IVR calling**
- **Multiple AI agents** running simultaneously
- **Access to 40+ AI agents** via multiple APIs, LiveKit, and ElevenLabs
- **Live dashboard to monitor calls,** hangups, call status on both TG and Web, plus after calls Transcript
- **Custom mixing of LLM, STT, and TTS** to match any use case
- **Configurable SIP trunk** directly from the bot
- **DTMF input collection based** on custom call flows
- **Telegram notifications and data capture** based on conversation context
- **Live call transfer to 3CX** or any custom SIP endpoint (can be customized for SIP)
- **Supports both inbound and outbound** calling

Performance

- Supports up to **100 concurrent calls** using a blended multi-LLM system (e.g., LiveKit up to 30, ElevenLabs up to 15, and more in parallel)

Figura 4: Anuncio de un bot de voz con IA para Telegram



NightRaider
Человек для всего
●●●●●

Active arbitrage
134
284 posts
Joined
02/20/25 (ID: 189648)
Activity
вирусология / malware
Deposit
0.006064 ₿
Autogarant
52

Posted April 9 (edited)

I'm selling the latest version of Cobalt Strike. No restrictions and unlimited usage. We are the only ones that can do it. Nobody else can.
Price: 7500$ and it's negotiable (Note that Cobalt Strike license is no longer 3500$. It has doubled)

Я продаю последнюю версию Cobalt Strike. Без ограничений и с неограниченным использованием. Мы единственные, кто может это сделать. Никто другой не может.
Цена: 7500$ и она обсуждаема (обратите внимание, что лицензия Cobalt Strike больше не стоит 3500$. Она подорожала вдвое)

If you already bought from me either BRC4 or Cobalt Strike, I'll offer a generous discount. Please contact me.
Если вы уже покупали у меня BRC4 или Cobalt Strike, я предложу вам щедрую скидку. Пожалуйста, свяжитесь со мной.

The new 4.12 update includes these changes:

GUI:

- Completely redesigned interface with theme support (Dracula, Solarized, Monokai, etc.)
- Improved Pivot Graph showing listener names and pivot types

REST API (Beta):

- Language-agnostic API to script CS from any language (Python, etc.)
- Task tracking with task/response relationships
- Central artifact management
- MCP server integration for Claude LLM compatibility

Figura 5: Un usuario de un foro de ciberdelincuencia anuncia una actualización de Cobalt Strike

Sin embargo, aunque existen pruebas claras de experimentación y adopción activas, la comunidad, al menos en los foros analizados por la CTU, no muestra un entusiasmo generalizado. En consonancia con las conclusiones anteriores de Sophos al analizar las actitudes hacia la IA generativa en foros delictivos en 2023 y 2025, la CTU observó perfiles que expresaban su preocupación por que la IA redujera las oportunidades de trabajo, especialmente en el desarrollo de malware y scripts. Otros rechazaban directamente la IA y recomendaban confiar en las capacidades humanas.

Ingeniería social y deepfakes mejorados mediante IA

La IA está modificando la rapidez con la que pueden ampliarse las campañas de ingeniería social, el grado de credibilidad que alcanzan en distintos idiomas y el bajo coste con el que pueden producirse. Por ejemplo, en el informe M-Trends 2026 de Mandiant se describía una transición desde el phishing genérico hacia una ingeniería social personalizada y orientada a crear vínculos mediante LLM. El phishing por voz ascendió hasta convertirse en el segundo vector de infección inicial más habitual, con un 11 %, mientras que el phishing por correo electrónico descendió hasta solo un 6 %. En el informe ConnectWise 2026 MSP Threat Report se confirmó el uso activo de fraudes mediante deepfakes y campañas de phishing generadas mediante LLM, y señaló que la IA había dotado de mayor rapidez, escalabilidad y credibilidad a tácticas conocidas, en lugar de crear nuevas categorías de ataques.

Los deepfakes se han convertido en una herramienta operativa para distintas categorías de ciberdelicuentes, especialmente los norcoreanos, cuyos operativos utilizan identidades generadas mediante IA y deepfakes para conseguir empleos remotos en empresas occidentales y establecer así un acceso interno persistente. Sophos ha publicado una guía específica para CISO sobre las tácticas de los trabajadores de TI norcoreanos que aborda los indicadores de detección y las medidas que pueden adoptar las organizaciones.

Sophos CTU investigó una estafa shā zhū pán que empleaba ingeniería social basada en IA. Una víctima del Reino Unido fue atraída hacia una plataforma de inversión falsa basada en IA llamada DAIQ Wealth mediante meses de "lecciones" sobre IA impartidas a través de la aplicación de mensajería LINE. Una red de perfiles estableció una relación de confianza mediante chats de grupo. Cada perfil estaba gestionado por una persona distinta, lo que apuntaba a la existencia de un equipo organizado que seguía guiones predefinidos. La víctima perdió cientos de miles de libras. Cuando expresó su preocupación por tener casi 20 000 GBP en efectivo en casa, los estafadores enviaron en un plazo de 24 horas a un transportista para recoger el dinero en su domicilio del Reino Unido, que incluso entregó un recibo con la marca de una empresa financiera legítima utilizado sin conocimiento de esta.

Resumen

Como ya hemos señalado, nuestro estudio de caso sobre STAC6994, la contratación clandestina de ingenieros de instrucciones, el uso de la IA como argumento comercial en servicios de ciberdelincuencia y las campañas fraudulentas relacionadas con la IA indican que los ciberdelicuentes están adoptando la IA para complementar y facilitar los ataques, a la vez que experimentan con ella.

En el conjunto del sector se observa un panorama similar. Claude facilitó ataques contra redes gubernamentales mexicanas. En el informe AI Risk and Resilience de Mandiant se describen dos familias de malware documentadas que consultan LLM durante su ejecución: PROMPTFLUX, un instalador experimental en VBScript que utiliza la API de Gemini para modificarse a sí mismo justo a tiempo; y PROMPTSTEAL, atribuido a IRON TWILIGHT (APT28), que constituye la primera observación confirmada de malware patrocinado por un Estado consultando un LLM durante operaciones reales contra Ucrania. En el DBIR de Verizon se documentó VoidLink, un marco de malware creado por un agente de IA en seis días, y PromptLock, un ransomware basado en IA descubierto por primera vez por ESET en agosto de 2025, aunque posteriormente se reveló que probablemente se trataba de una prueba de concepto académica y no de una muestra maliciosa detectada en circulación.

Sin embargo, las campañas de intrusión totalmente autónomas dirigidas por IA siguen sin confirmarse a gran escala. En el informe sobre ransomware GRIT 2026 de GuidePoint se afirmó expresamente que se ha exagerado la preocupación por los "superafiliados que despliegan bots de ransomware totalmente autónomos" y que el uso de IA y LLM continúa siendo rudimentario entre los ciberdelicuentes menos avanzados. La evaluación AIM3 de Recorded Future situó la mayor parte del malware con IA observado en los niveles de madurez del 1 al 3 de su modelo de madurez del malware con IA: experimentación, adopción y optimización. También señaló que no existen ejemplos confirmados de malware con IA propia verdaderamente integrada que ejecute modelos locales en los hosts de las víctimas y concluyó que "los responsables de la seguridad deben priorizar la supervisión del uso indebido de servicios de IA legítimos, reforzar los controles existentes y asociar las amenazas con los niveles de AIM3, en lugar de reaccionar de forma desproporcionada ante escenarios de ciencia ficción".

El caso GTG-1002 documentado por Anthropic, en el que Claude Code automatizó entre el 80 % y el 90 % de una campaña de robo de datos contra aproximadamente 30 entidades, es el caso de funcionamiento autónomo más cercano que consta en registros públicos, pero sigue siendo una anomalía y no un ejemplo representativo de una tendencia.

Parte 2:
Riesgos de la IA empresarial

La IA ya está presente en las empresas: los agentes de programación escriben y despliegan código con credenciales de desarrollador, los asistentes agénticos leen correos electrónicos y llaman a las API, se conectan entre sí varios sistemas y los equipos internos experimentan con modelos de pesos abiertos en infraestructuras de nube gestionadas (los modelos de código abierto y pesos abiertos son cada vez más capaces y suelen costar una fracción de lo que cuestan los modelos de código cerrado y pesos cerrados. Además, permiten que las empresas sean propietarias de sus inferencias y conserven el control de sus datos, aunque proceden en gran medida de la República Popular China). La cuestión de gobernanza ya no es si debe permitirse su adopción, sino cómo protegerla antes de que la superficie de ataque quede consolidada en torno a configuraciones predeterminadas deficientes.

Incluso las organizaciones más reacias al riesgo que no utilizan ampliamente la IA en casos de uso autorizados probablemente tengan un problema de IA en la sombra y, como veremos en breve, los temores relacionados con la IA en la sombra y las fugas de datos están fundamentados. En la parte 4 analizaremos en detalle la tolerancia al riesgo, los controles y la visibilidad.

Una tendencia diferenciada en 2026 es el ataque contra infraestructuras específicas de desarrollo de IA y, en particular, contra las relaciones de confianza que los desarrolladores establecen en torno a las herramientas de IA.

La infraestructura de desarrollo de IA como objetivo de los ataques

Una tendencia diferenciada en 2026 es el ataque contra infraestructuras específicas de desarrollo de IA y, en particular, contra las relaciones de confianza que los desarrolladores establecen en torno a las herramientas de IA. A diferencia de la mayoría de las amenazas relacionadas con la IA, que siguen siendo emergentes o se limitan en gran medida a pruebas de concepto teóricas, esta es un área en la que los ataques ya se están produciendo.

La campaña del gusano npm **SANDWORM_MODE** implicó al menos 19 paquetes de typosquatting diseñados para imitar herramientas legítimas para desarrolladores y herramientas de programación con IA. Los paquetes instalaban un servidor MCP no autorizado y, mediante una inyección de instrucciones integrada, obligaban a asistentes de IA legítimos a recuperar de forma encubierta claves SSH y credenciales en la nube y a exfiltrarlas sin que el usuario lo supiera.

La **vulneración de la extensión Nx Console para VS Code** recopiló credenciales de HashiCorp Vault, npm, AWS, GitHub y 1Password, así como claves de API de Anthropic, y parecía formar parte de la campaña más amplia TeamPCP / mini-Shai-Hulud. Una extensión vulnerada dentro del IDE de un desarrollador tiene acceso directo a las credenciales y los repositorios más importantes.

La **filtración del código fuente de Anthropic Claude Code** en marzo de 2026, cuando el código fuente se publicó por error en un paquete npm, ilustra una dimensión diferente. Según la información publicada, el código contenía aproximadamente 1900 archivos y 500 000 líneas de código, incluidos detalles de funciones aún no publicadas. Aunque Anthropic afirmó que no se expusieron datos de clientes, el análisis del código podría permitir identificar errores y vulnerabilidades en el futuro. Las propias herramientas de IA se han convertido en un objetivo para la recopilación de información.

Nuestra recomendación: cualquier organización que cuente con un equipo de desarrollo debe considerar este su principal riesgo inmediato. Los controles en este ámbito dependen en gran medida de los procesos, como una gestión cuidadosa de las dependencias, la fijación de versiones y la revisión de las nuevas bibliotecas de código abierto antes de incorporarlas, junto con una supervisión estrecha de los endpoints de los desarrolladores (en la parte 4 ofrecemos una lista de siete medidas que los responsables de la seguridad pueden adoptar ahora para reducir el alcance de los posibles daños derivados del despliegue de IA agéntica).

Dónde se concentra la exposición

El entramado de identidades que conecta los servicios de IA con los sistemas empresariales genera una exposición para la que la gobernanza existente no fue diseñada. IBM X-Force afirmó que durante 2025 se pusieron a la venta en la Web Oscura más de 300 000 credenciales de ChatGPT y que, en una publicación de un foro de febrero de 2025, se aseguraba que se habían robado más de 20 millones de cuentas de ChatGPT, aunque esta cifra continúa sin confirmarse. El incidente de Salesloft/Drift demostró un mecanismo concreto: se utilizaron tokens de OAuth de Drift para vulnerar varios entornos de Salesforce, lo que demuestra que el acceso basado en tokens de un chatbot puede traducirse directamente en la vulneración de un sistema.

El entramado de identidades que conecta los servicios de IA con los sistemas empresariales genera una exposición para la que la gobernanza existente no fue diseñada.

BeyondTrust informó de un aumento del 466,7 % en el número de agentes de IA activos dentro de los entornos empresariales durante el último año. Esos agentes introducen superficies de ataque específicas: inyección de instrucciones, invocación de herramientas maliciosas y entradas de datos envenenadas. La vulnerabilidad CVE-2025-32711 (EchoLeak) del propio Copilot de Microsoft demostró cómo una depuración insuficiente de las entradas puede permitir fugas de datos sin necesidad de interacción. Cuando las identidades de máquina carecen de equivalentes a la MFA, utilizan claves de larga duración y operan con privilegios excesivos, se convierten en un objetivo viable y atractivo.

Los atacantes son plenamente conscientes de estas oportunidades y están identificando activamente la superficie de ataque. En enero de 2026, GreyNoise describió campañas destinadas a mapear los despliegues de LLM. Los investigadores documentaron la Operación Bizarre Bazaar: una campaña que secuestraba endpoints de LLM y MCP expuestos, los validaba y revendía el acceso mediante una infraestructura vinculada a silver.inc.

La laguna de gobernanza

En todo el sector, y especialmente entre la alta dirección, existe preocupación por las implicaciones de la IA para la seguridad, así como una falta de capacidad y recursos para abordarlas.

Solo ChatGPT generó 410 millones de infracciones de políticas DLP en los datos de Zscaler. En el informe Saviynt/Cybersecurity Insiders CISO AI Risk Report 2026 se reveló que el 75 % de los encuestados había descubierto herramientas de IA en la sombra y que el 95 % dudaba de su capacidad para detectar o contener su uso indebido. Solo el 1 % de las empresas dispone de un presupuesto específico para la seguridad de la IA (Pentera) y, según el informe 2026 CISO Report de Splunk, aunque existe consenso en que la IA puede contribuir a mejorar la productividad y gestionar el volumen de trabajo, a la mayoría de los CISO les preocupan las fugas de datos, la IA en la sombra y las alucinaciones. Curiosamente, Splunk constató que estos temores eran mayores entre los CISO de organizaciones con una mayor madurez en IA y sugirió que quienes "se encuentran en una fase más avanzada de su andadura con la IA generativa están empezando a percibir algunos inconvenientes".

Estos inconvenientes y preocupaciones, especialmente los relacionados con la IA en la sombra, están fundamentados. En abril de 2026, Vercel comunicó una vulneración provocada por el uso que hizo un empleado de Context.ai, una herramienta de productividad de IA de terceros. El empleado se registró con su cuenta corporativa de Google Workspace y concedió permisos de "Permitir todo" a la aplicación OAuth de Context.ai. Cuando Context.ai fue vulnerada, al parecer mediante una infección de Lumma Stealer en el dispositivo de un empleado de Context.ai, los atacantes obtuvieron esos tokens de OAuth y los utilizaron para acceder a los entornos internos de Vercel.

El reto de la gobernanza reside en que la adopción de la IA avanza más rápido que los procesos organizativos diseñados para gestionarla. Además, el entorno normativo se está endureciendo, mientras aumenta la brecha entre la velocidad de despliegue y la madurez de la gobernanza.
La Ley de IA de la UE exige que los sistemas de alto riesgo la cumplan plenamente en 2026, con sanciones de hasta el 7 % de la facturación mundial. Los requisitos de transparencia de California entraron en vigor el 1 de enero de 2026. Sin embargo, Saviynt/Cybersecurity Insiders determinaron que, aunque el 71 % de las grandes empresas ha desplegado agentes de IA con acceso a sistemas empresariales esenciales, solo el 16 % gobierna eficazmente dicho acceso.

Como punto de partida para abordar esta brecha, el tráfico de instrucciones debe tratarse como un registro de seguridad de primer nivel: la captura y el análisis de instrucciones pueden revelar intentos de inyección, exfiltración, uso indebido por parte de usuarios internos y comportamientos inesperados de los agentes. Esto puede comenzar en la capa de proxy para los modelos alojados en la nube o en la capa de orquestación para los agentes desplegados localmente, sin tener que esperar a que se complete la integración del modelo. Pero los instrucciones solo son una parte de la respuesta; las herramientas y las habilidades se están convirtiendo en una parte fundamental de la superficie de ataque y deben tratarse como tal.

La adopción de la IA avanza más rápido que los procesos organizativos diseñados para gestionarla y el entorno normativo se está endureciendo

Parte 3:
Cómo está cambiando la IA los modelos operativos de defensa

Clasificación e investigación

La ejecución de estrategias, la reconstrucción de líneas de tiempo y la elaboración de informes en lenguaje natural son ámbitos en los que los modelos de razonamiento pueden aportar un verdadero valor operativo. Por ejemplo, la infraestructura de MDR de Kaspersky procesó aproximadamente 400 000 alertas en 2025, y una lógica de detección basada en IA filtró los falsos positivos antes de que los revisaran los analistas; 39 000 alertas se derivaron para una investigación más exhaustiva. Los modelos de IA tiene cada vez mayor capacidad para llevar a cabo esas investigaciones; en marzo de 2026, Prophet Security destacó en The Hacker News dos casos en los que los modelos realizaron análisis contextuales e investigaciones de incidentes cuya actividad maliciosa no resultaba evidente de inmediato. El primero implicaba a un usuario inactivo con una antigua clave de API que se activó repentinamente; el segundo consistía en descubrir la intención maliciosa de un correo electrónico de phishing que habría superado la mayoría de las comprobaciones de seguridad y validación, si no todas. Prophet Security señala que, en ambos casos, ningún dato aislado resultaba sospechoso por sí mismo; la amenaza solo se hizo evidente mediante un análisis contextual de múltiples datos, el tipo de análisis que realizan los investigadores humanos cuando disponen de capacidad para ello. El problema es, sin duda, que los investigadores a menudo no disponen de esa capacidad, y ahí es donde pueden contribuir los modelos de razonamiento, ya que proporcionan el mismo nivel de análisis a todas las alertas.

Sin embargo, la IA no es una panacea y requiere un diseño meticuloso. Los estudios sobre el razonamiento excesivo demuestran que, a partir de cierto umbral, un razonamiento adicional hace que los modelos cambien respuestas correctas por incorrectas. Los cambios de respuestas correctas a incorrectas superaron a los cambios en sentido contrario después de aproximadamente 7000 tokens, hasta alcanzar una proporción de tres a uno al llegar a los 12 000. Por tanto, las implementaciones del SOC deben considerar el esfuerzo de razonamiento, la profundidad de los reintentos y los bucles de llamadas a herramientas como parámetros de ingeniería ajustables.

En términos generales, la adopción es desigual y es necesario ajustar las expectativas. Según el informe de Splunk sobre los CISO, actualmente solo el 40 % de los CISO utiliza IA generativa en sus funciones de seguridad. La IA agéntica solo registra un 6 % de despliegue activo, aunque el 39 % está explorando su uso.

También hay que tener en cuenta que las herramientas de IA destinadas a los responsables de la seguridad pueden tener su propia superficie de riesgo. Por ejemplo, en el informe State of Pentesting de Cobalt se reveló que las aplicaciones de IA y LLM presentan hallazgos de alto riesgo con una frecuencia 2,7 veces superior a la del software tradicional, y que solo el 38 % de los problemas se resolvió durante las pruebas de penetración.

El problema de la memoria

La mayoría de los despliegues de producción permanecen estancados en lo que algunos investigadores describen como la segunda o tercera generación de madurez agéntica: asistentes que utilizan herramientas y comienzan desde cero en cada sesión. Conservan su entrenamiento, pero no pueden recordar que el mismo patrón de alerta apareció el día anterior y resultó ser benigno, ni que el bloqueo de un intervalo de direcciones IP concreto interrumpió la VPN el mes anterior.

La diferencia entre un analista sénior y uno júnior no radica necesariamente en su capacidad de razonamiento, sino en lo que aportan a cada evento: memoria episódica de incidentes anteriores, conocimiento semántico del entorno e intuiciones procedimentales sobre lo que funciona. Los asistentes de IA actuales se parecen más a analistas júnior que sénior e investigan cada alerta como si fuera la primera vez. Los investigadores están estudiando arquitecturas de memoria como sistema operativo, servicios externos de memoria como middleware y enfoques basados en grafos de conocimiento con razonamiento bitemporal. Hasta que estas tecnologías maduren, los sistemas de IA seguirán siendo asistentes útiles, en lugar de operadores autónomos.

El límite de la detección

Los sistemas de detección de comportamientos procesan flujos de datos de telemetría en directo con un desequilibrio extremo entre clases. Una plataforma que procesa billones de eventos diarios los reduce a millones de detecciones deduplicadas, prioriza miles de alertas de gravedad alta y genera cientos de investigaciones diarias que siguen considerándose críticas tras la revisión humana. La proporción integral se aproxima a uno entre diez mil millones. El trabajo fundamental de Stefan Axelsson sobre la falacia de la tasa base en la detección de intrusiones formalizó las matemáticas: incluso un detector con una especificidad del 99,9 % genera millones de falsas alarmas por cada amenaza real cuando la probabilidad previa de ataque es lo suficientemente baja.

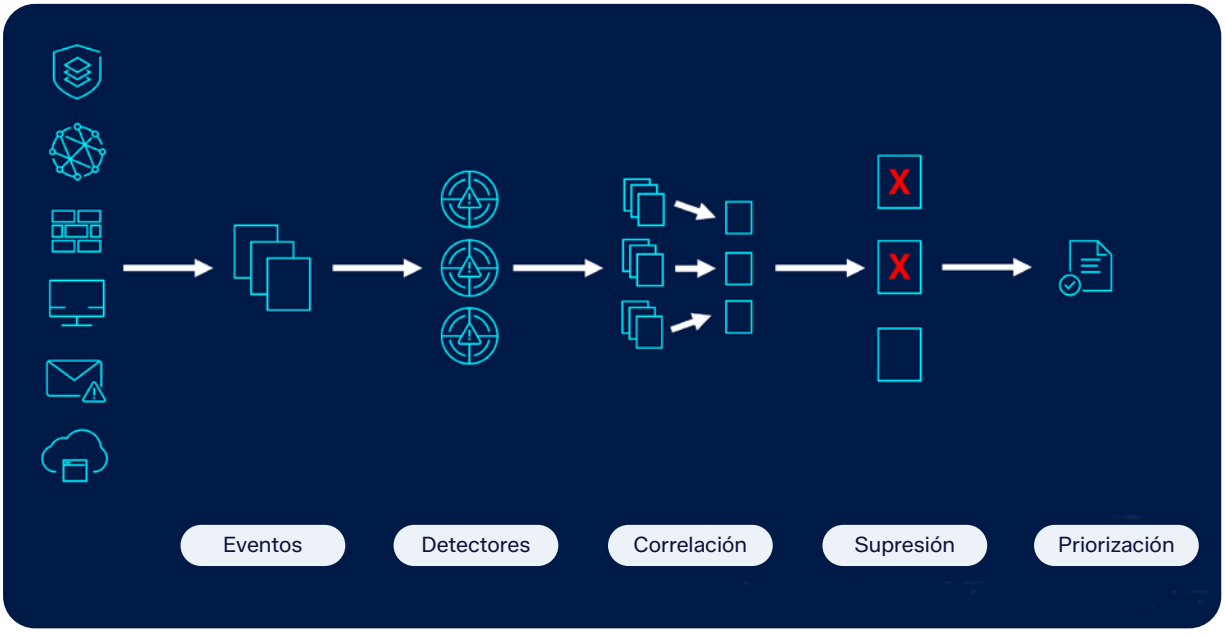


Figura 6: Descripción general del proceso de varias fases

La investigación de Sophos presentada en NorthSec 2026 demuestra por qué la detección por capas sigue siendo importante. Los datos de telemetría recopilados durante un periodo de dos semanas, que comprendían 11,8 billones de eventos, se redujeron mediante detectores de complejidad creciente, como la correspondencia sencilla de indicadores, las reglas de correlación, los modelos de Machine Learning específicos, la deduplicación, la correlación y la supresión, hasta quedar en 81 573 alertas de gravedad alta y crítica, lo que supone una media inferior a 50 por organización. Entre esas alertas restantes, Sophos identificó un ataque de infostealer relacionado con la campaña TamperedChef.

Los modelos de razonamiento deben situarse por encima de esta capa. Generan hipótesis de investigación, recuperan y sintetizan datos de telemetría para producir resúmenes de pruebas y coordinan acciones de respuesta limitadas mediante protecciones explícitas. No sustituyen a los modelos entrenados con datos propios de incidentes, referencias específicas de cada entorno y bucles de aprendizaje continuo a los que los sistemas de razonamiento general no tienen acceso.

En la Encuesta de Madurez de Confianza en IA de McKinsey cuantificó la limitación organizativa: las cuestiones de seguridad constituyen el principal obstáculo para ampliar el uso de la IA agéntica para casi dos tercios de los encuestados. La falta de precisión (74 %) y la ciberseguridad (72 %) son los principales riesgos específicos. Solo alrededor del 30 % de las organizaciones ha alcanzado el nivel de madurez tres ("Se están adoptando medidas para desarrollar todas las prácticas necesarias de IA responsable"), o un nivel superior en la gobernanza de la IA. Las organizaciones que han demostrado explícitamente la propiedad de la IA responsable alcanzan una madurez media de 2,6, frente al 1,8 de aquellas que no han definido claramente dicha responsabilidad.

La preparación de la organización es el factor que limita la defensa asistida por IA. Los modelos cuentan con capacidades suficientes; la cuestión es si las organizaciones pueden utilizarlos con el nivel de confianza que requieren unas operaciones de seguridad autónomas.

Investigación de Sophos AI: Defensa de la capa de IA

Además de la investigación sobre la detección por capas descrita anteriormente, el equipo de Sophos AI ha desarrollado otras técnicas de defensa novedosas, cada una de ellas destinada a abordar un modo de fallo específico.

- El **Untrusted Data Scanner (UDS)** inspecciona el contenido externo para detectar inyecciones de instrucciones antes de que llegue a un LLM o un agente. Todos los casos de uso de LLM tienen un canal de instrucciones, que contiene lo que la aplicación indica al modelo que debe hacer, y un canal de datos, que contiene el contenido no fiable que procesa. La inyección de instrucciones introduce instrucciones de forma encubierta en el canal de datos. El UDS está deliberadamente limitado a este problema específico, en lugar de actuar como detector general de jailbreaks. Los primeros resultados muestran un índice reducido de falsos positivos, algo importante porque los datos de seguridad están repletos de lenguaje relacionado con la seguridad, como informes de incidentes, análisis de malware y hallazgos de pruebas de penetración, y un detector que generara una falsa alarma con cada informe de amenazas resultaría inutilizable.
- **LLM Scoping** clasifica si las consultas se encuentran dentro de la función prevista de una característica y rechaza las solicitudes que quedan fuera de su ámbito antes de que lleguen al LLM. En los métodos evaluados, los modelos de delimitación del alcance superaron de forma considerable a las defensas basadas en instrucciones del sistema y redujeron prácticamente a cero los índices de éxito de los ataques en varias pruebas comparativas públicas. El equipo también aplicó Multi-round Automatic Red Teaming (MART) para reforzar los modelos de delimitación del ámbito frente a la adaptación de los adversarios.

- **CerBERTus**, un modelo basado en BERT con "tres cabezas", aborda la fragilidad de la detección binaria de jailbreaks. Un único codificador compartido alimenta tres cabezas de clasificación: nivel de perjuicio, como función principal; categoría del objetivo, es decir, qué intenta hacer el usuario; y estilo de formulación, es decir, cómo se presenta la solicitud. Las tareas auxiliares actúan como sesgo inductivo y hacen que la representación diferencie el objetivo del contenedor. El equipo creó un corpus factorial estructurado de instrucciones que combina objetivos, como ciberataques, fraude, explosivos, síntesis de drogas, privacidad o propaganda extremista, entre otros, con marcos de adversarios, como juegos de roles, ficción, urgencia, pretextos académicos u ofuscación, para entrenar y someter a pruebas de resistencia esta separación.
- **Detección de anomalías**: En una investigación presentada en Black Hat USA 2025, el equipo combinó la detección de anomalías con LLM y descubrió que el éxito del método procedía de la generación de líneas de comandos benignas y diversas, no de la localización de líneas maliciosas, lo que redujo considerablemente los índices de falsos positivos en los clasificadores de líneas de comandos.
- El **salado (salting) de LLM**, presentado en CAMLIS 2025, se inspira en el salado de contraseñas: introduce variaciones de comportamiento específicas para impedir la reutilización de jailbreaks precalculados en despliegues homogéneos de LLM.

Engaño frente a atacantes autónomos

Durante dos décadas, el engaño cibernético nunca llegó a convertirse en un control principal. Frente a atacantes autónomos, el planteamiento cambia. Los agentes de IA no se cansan y pueden enumerar todas las rutas a velocidad de máquina. Los entornos engañosos, como los propuestos en el cebo para agentes LLM de Palisade y el marco MANTIS, pueden llenar el contexto de un agente con información engañosa, agotar su presupuesto de inferencia e imponer costes económicos reales. Un ejemplo de ello, HoneyTrap, demostró un aumento del 149 % en el "consumo de recursos del atacante" en comparación con las defensas convencionales. Este ámbito aún se encuentra en una fase inicial, pero las condiciones en las que el engaño se convierte en una defensa principal son precisamente las que está creando la ampliación de la IA ofensiva.

Dónde repercute la asimetría del ritmo

Un atacante que utiliza IA no se enfrenta a ningún ciclo de adquisición, revisión de cumplimiento ni comité asesor de cambios. Un ciberdelicuente puede adoptar una herramienta y aplicar sus resultados en cuestión de días. La campaña asistida por IA contra dispositivos FortiGate vulneró más de 600 firewalls en 55 países en cinco semanas y fue obra de un ciberdelicuente que utilizaba herramientas comerciales de IA.

Por parte de los responsables de la seguridad, se activa una detección y la investigación asistida por IA genera una recomendación de contención en cuestión de minutos, pero el proceso de remediación del cliente requiere una solicitud de cambio, una ventana de mantenimiento y un acuerdo de nivel de servicio (SLA) de 48 horas La investigación se aceleró, pero la respuesta no. Esto fue posiblemente lo que ocurrió con dos vulnerabilidades explotadas rápidamente durante el último año. Como señalamos anteriormente, CVE-2026-10520 (Ivanti Sentry) se explotó en las 24 horas posteriores a la publicación de la prueba de concepto y CVE-2026-42208 (LiteLLM), en un plazo de 36 horas.

Aunque la clasificación asistida por IA ayuda sin duda a los responsables de la seguridad a mantener el ritmo, el trabajo previo en los principios fundamentales de seguridad puede contribuir a corregir esta asimetría: reducir la superficie de ataque, aplicar la MFA y mantener unos datos de telemetría lo suficientemente completos como para reconstruir lo sucedido cuando falla la prevención.

Parte 4:
Qué deben hacer los responsables de seguridad

Las pruebas incluidas en este informe respaldan un reducido número de decisiones condicionadas por la situación actual de cada organización.

Calibrar la tolerancia al riesgo para adoptar la IA

Las pruebas incluidas en este informe generan una tensión que las organizaciones deben resolver: la adopción de la IA conlleva riesgos reales, pero una adopción lenta también. Avanzar con demasiada cautela mientras los atacantes aceleran implica quedarse atrás en el ritmo defensivo, que este informe identifica como el reto determinante. Aceptar riesgos forma parte inherente de este proceso; la cuestión es cómo asumir los riesgos adecuados.

Las pruebas incluidas en este informe generan una tensión que las organizaciones deben resolver: la adopción de la IA conlleva riesgos reales, pero una adopción lenta también.

El marco interno de riesgos de Sophos diferencia los riesgos adecuados de los inadecuados según dos ejes: el potencial de beneficio y la contención de las consecuencias negativas. Un riesgo adecuado tiene un impacto significativo, medidas para contener el alcance de los posibles daños, una vía reversible, ya sea mediante un proyecto piloto, una reversión o una alternativa, una responsabilidad claramente asignada y ciclos de comentarios definidos. Un riesgo inadecuado presenta un potencial de daño ilimitado, incumple aspectos no negociables como la seguridad, la privacidad, los requisitos legales o el cumplimiento, carece de un plan de recuperación, no tiene una responsabilidad claramente asignada u omite el criterio humano en cambios que afectan directamente a los clientes.

En el caso concreto de la IA, el marco se traduce en una serie de preguntas prácticas antes de cualquier despliegue: ¿cuál es el peor resultado posible y verosímil? ¿Cuál es el posible alcance de los daños? ¿Podemos revertirlo? ¿Quién es el propietario? Los despliegues cuyas consecuencias negativas puedan contenerse, como un proyecto piloto de solo lectura de una herramienta de clasificación agéntica sobre un conjunto de datos aislado, deben llevarse a cabo con decisión. Los despliegues con consecuencias negativas importantes, como conceder a un agente acceso de escritura a la infraestructura de producción, deben someterse a medidas de reducción del riesgo o a límites de alcance antes de continuar. Los despliegues con poco impacto y consecuencias negativas importantes deben evitarse por completo.

Los aspectos no negociables se aplican en cualquier circunstancia: confianza del cliente, protección, seguridad y privacidad, requisitos legales y cumplimiento, y estabilidad operativa. La función del CISO en un entorno acelerado por IA consiste en garantizar que la organización asuma riesgos inteligentes y evite los imprudentes.

Para las organizaciones que despliegan IA agéntica: reducción del alcance de los posibles daños

Sophos ha propuesto siete controles, aplicables en un plazo de entre uno y seis meses, para reducir el alcance de los posibles daños en las organizaciones que desean desplegar IA agéntica:

1. **Los espacios seguros de los agentes** establecen un límite controlado alrededor del proceso del agente. La mayoría de los entornos de ejecución incluyen algún tipo de espacio seguro, pero normalmente es opcional. Siempre que sea posible, elija entornos de ejecución remotos o basados en la nube en lugar de entornos aislados locales para conseguir una separación más sólida de los procesos.
2. **El aislamiento de credenciales** garantiza que el agente nunca vea directamente los secretos: un proceso independiente obtiene las credenciales de una caja fuerte, las introduce en la llamada a la API y devuelve únicamente respuestas depuradas, manteniendo las credenciales de larga duración fuera de las ventanas de contexto del LLM.
3. **Los endpoints de herramientas aislados** van más allá. El agente llama a una herramienta por su nombre, pero un proceso intermediario conserva la credencial, realiza la llamada real a la API conforme a un esquema fijo y aplica listas de destinos de salida permitidos para cada herramienta. La única capacidad de decisión del agente consiste en elegir qué herramienta invocar y qué parámetros proporcionar.
4. **La restricción del tráfico saliente y la monitorización de la red** consideran al agente vulnerado como un canal de exfiltración de datos: detección de secretos en el tráfico saliente, umbrales de volumen para las solicitudes salientes y clasificación web para bloquear las solicitudes a dominios sin categorizar o registrados recientemente.
5. **La cobertura de EDR** debe extenderse al comportamiento del host del agente: los contenedores, las máquinas virtuales efímeras y los equipos no administrados de los desarrolladores son puntos ciegos habituales.
6. **La aprobación con intervención humana** separa el plano de ejecución autónoma del agente de un plano de control gobernado por personas: las operaciones irreversibles requieren un procedimiento criptográfico, en lugar de un simple botón de confirmación que el agente pudiera simular.
7. **Los límites de propagación de las inyecciones** abordan el riesgo creciente a medida que las inyecciones pasan de estar limitadas a una sesión a extenderse al agente y, posteriormente, a propagarse entre agentes. Trate las escrituras en memoria y las transferencias entre agentes como eventos de seguridad.

Cuando OpenClaw ganó popularidad a principios de 2026, más de 30 000 instancias quedaron expuestas en Internet y los ciberdelicuentes ya debatían cómo convertir sus "habilidades" modulares en armas para campañas de red de bots. El Red Team de Sophos lo puso a prueba directamente: 23 hallazgos procesables, un reconocimiento de Active Directory reducido de tres días a tres horas y un nivel de detalle en el registro de auditoría que no habría sido posible conseguir manualmente en un plazo razonable. Sin embargo, el equipo dedicó más tiempo a establecer barreras de seguridad que a ejecutar la prueba. La conclusión: la IA agéntica es lo suficientemente potente como para que merezca la pena desplegarla, pero solo si se establece primero el marco de seguridad.

Visibilidad, identidad y aplicación de parches

Recomendamos comenzar por obtener visibilidad sobre el uso de los servicios de IA. La exposición relacionada con la IA documentada en los incidentes de Salesloft/Drift y Vercel convierte esta medida en el punto de partida de mayor impacto. Haga un inventario de las herramientas de IA de su entorno. Identifique cuáles de ellas almacenan tokens de OAuth, claves de API o cuentas de servicio conectadas a sistemas empresariales. Determine qué datos circulan a través de ellas y de qué permisos disponen. No puede gobernar aquello que no puede ver.

Este trabajo de visibilidad debe dar lugar a un registro de IA que incluya todas las herramientas, agentes y modelos de IA, las credenciales a las que pueden acceder, la configuración de espacios seguros y los requisitos de conectividad previstos. Sin este registro, no pueden aplicarse sistemáticamente controles como el aislamiento de credenciales, la restricción del tráfico saliente y los espacios seguros. Aplique controles de identidad con el mismo rigor que a cualquier otra aplicación SaaS: aplique políticas de acceso condicional, exija la MFA para las cuentas de servicio de IA, aplique controles DLP a las rutas de salida que utilizan las herramientas de IA y rote las credenciales de las integraciones conectadas a la IA con la misma periodicidad que cualquier otro acceso privilegiado. Tanto las conclusiones de IBM X-Force sobre los datos de credenciales como el ataque a la cadena de suministro de Nx Console confirman que las credenciales de los servicios de IA son un objetivo de recopilación.

La explotación asistida por IA está reduciendo a horas el intervalo entre la divulgación y la explotación.

Las organizaciones cuya periodicidad de aplicación de parches aún se mide en semanas se enfrentan a un riesgo creciente. El plazo de tres días establecido por la directiva BOD 26-04 de la CISA constituye el mínimo, ya que la explotación asistida por IA está reduciendo a horas el intervalo entre la divulgación y la explotación. Si una vulnerabilidad crítica expuesta a Internet no puede corregirse en cuestión de días, esa laguna conlleva un riesgo cuantificable.

Evaluar con pruebas

Instrumente las herramientas de IA que utilizan sus analistas. Mida el esfuerzo de validación, es decir, cuánto tiempo dedican los analistas a comprobar los resultados de la IA. Mida la eficiencia de los instrucciones, es decir, cuántos intentos se necesitan para obtener resultados útiles. Y mida la calibración de la confianza, es decir, si los analistas confían demasiado o demasiado poco en el sistema. Cree conjuntos de pruebas a partir de cargas de trabajo reales del SOC, en lugar de pruebas comparativas genéricas, y controle las versiones de sus procedimientos para que sigan siendo válidos tras las actualizaciones de los modelos.

Proteger la cadena de suministro de la IA

La IA introduce riesgos para la cadena de suministro que van más allá de las dependencias de software tradicionales. Los pesos de los modelos, la procedencia de los datos de entrenamiento, la integridad de los servidores MCP y la propia infraestructura de inferencia constituyen límites de confianza. Las organizaciones que despliegan modelos de pesos abiertos mediante proveedores de servicios gestionados en la nube deben confirmar la versión exacta del modelo, la región desde la que se presta el servicio, el periodo de retención, la política de uso para el entrenamiento, la superficie de registro y las condiciones contractuales antes de enviar código o datos privados. En el caso de los modelos prestados mediante endpoints SaaS con sede en la República Popular China, como la API de DeepSeek, debe aplicarse un límite más estricto: estos modelos deben utilizarse para probar capacidades con datos públicos o sintéticos, no con repositorios propios ni con información confidencial de ingeniería.

Conclusión

El marco de los ciberdelicuentes de STAC6994 contenía perfiles de Cobalt Strike, módulos para eludir la EDR, herramientas para recopilación de credenciales, utilidades de movimiento lateral y un canal de comando y control (C2) oculto tras una infraestructura legítima en la nube. Todos los componentes nos resultaban conocidos, excepto el proceso de desarrollo. Los agentes de IA iteraron sobre técnicas de evasión más rápidamente de lo que habría podido hacerlo un desarrollador humano, las probaron contra productos específicos en un laboratorio estructurado, controlaron las versiones de los resultados mediante Git y entregaron el resultado a un operador que desplegó ransomware y robó datos.

En los casos de Sophos MDR e IR, la información de la CTU, los análisis de SophosLabs, la investigación de Sophos AI y el sector de la seguridad en general, la IA está acelerando operaciones conocidas en ambos lados de la seguridad. Ayuda a los atacantes a avanzar más rápidamente por cadenas de ataque conocidas. Permite a los responsables de la seguridad clasificar e investigar en plazos más reducidos. Crea nuevas exposiciones a través de la capa de identidad y gobernanza relacionada con la adopción de la IA para empresas. No ha producido un tipo de intrusión fundamentalmente nuevo que las arquitecturas de detección existentes no puedan abordar. El Instituto de Seguridad de la IA de Reino Unido constató que la capacidad ofensiva autónoma se había multiplicado aproximadamente por seis en 18 meses. Cada nueva generación de modelos permitió completar fases de escenarios de intrusión de varias fases que las generaciones anteriores no podían ejecutar. Esta trayectoria no permite caer en la complacencia.

Tres temas refuerzan la conclusión:

1. **La clasificación** debe ser precisa, porque agrupar todas las amenazas basadas en IA oculta la estrategia de defensa que requiere cada una.
2. **La evaluación** debe ser rigurosa, porque, tanto si se trata de la asignación de modelos como de la supresión de alertas o de la profundidad de razonamiento de los agentes, la respuesta debe obtenerse mediante mediciones.
3. La **transparencia** debe ser real, porque la confianza se deposita en las organizaciones que muestran su trabajo y comparten los fallos útiles junto con los éxitos.

Aspectos que Sophos observará durante los próximos 12 meses: si los agentes autónomos pasan de ejecutar estrategias conocidas para descubrir nuevas rutas de ataque a gran escala; si la economía sumergida desarrolla herramientas ofensivas de IA comercializadas como productos que reduzcan las barreras de entrada de forma tan drástica como el ransomware como servicio las redujo para la extorsión; si la gobernanza de la IA para empresas madura con la suficiente rapidez como para subsanar la exposición de la identidad creada por la actual oleada de adopción; y si el intervalo entre la explotación y la aplicación de parches continúa reduciéndose, obligando a replantearse la situación de las organizaciones que todavía miden su ritmo de respuesta en semanas.

El tiempo ha cambiado. Los responsables de la seguridad que no adapten su ritmo se encontrarán más rezagados de lo que suponían.

Colaboradores

Nash Borges
Vicepresidente sénior de Ingeniería e IA

Adarsh Kyadige
Responsable sénior de Investigación en IA

Ross McKerchar
CISO

Rafe Pilling
Director de Información sobre amenazas
Counter Threat Unit de X-Ops

Matt Stec
Director
Sophos AI

Ryan Westman
Responsable sénior de Investigación de amenazas
X-Ops Insights

Matt Wixey
Investigador sénior de amenazas
X-Ops Insights

Nuestro equipo

Sophos X-Ops es una iniciativa de ciberseguridad de vanguardia que reúne a más de 1000 expertos de diversos ámbitos especializados en dominios de seguridad dentro de Sophos, como información sobre amenazas, inteligencia artificial, operaciones de MDR y operaciones de seguridad internas.

Este equipo multidisciplinar refuerza las defensas de las organizaciones frente a las ciberamenazas actuales, altamente sofisticadas y en rápida evolución.

AI aprovechar los conocimientos combinados de sus expertos, Sophos X-Ops ofrece una respuesta multidimensional frente a los ciberataques y garantiza capacidades integrales de protección, detección y respuesta.

Este enfoque colaborativo e innovador ofrece una respuesta a amenazas sin precedentes y sitúa a Sophos como referente de excelencia y líder en ciberseguridad.

Sophos X-Ops aprovecha los conocimientos combinados de su equipo multidisciplinar. La sinergia entre los equipos multidisciplinares de Sophos X-Ops fomenta la información compartida y les permite adaptarse rápidamente a nuevos hallazgos, acelerando los tiempos de detección y respuesta, y reforzando las capacidades generales de protección para los clientes de Sophos.

SOPHOS

Para hablar sobre sus necesidades de seguridad de la IA y descubrir cómo puede ayudarle Sophos, visite nuestro sitio web o hable con un asesor.

Ventas en España
Teléfono: (+34) 913 756 756
Correo electrónico: comercialES@sophos.com

Ventas en América Latina
Correo electrónico: Latamsales@sophos.com

2026-07-20 ES (NP) CRE-5820