Seguridad para la inteligencia artificial: el nuevo frente crítico

Seguridad para la inteligencia artificial: el nuevo frente crítico

La explosión de la inteligencia artificial generativa ha aumentado la superficie de exposición. Ataques adversariales, manipulación de prompts, robo de modelos o fugas de datos son ya amenazas que se han vuelto reales. Frente a este escenario, proteger la IA —y protegerse con IA— se convierte en el nuevo imperativo para los equipos de seguridad.

La inteligencia artificial ha pasado en pocos años de ser una tecnología emergente a ocupar un lugar central en la estrategia de muchas organizaciones. No cabe duda de que su aplicación en entornos corporativos está transformando áreas como el análisis de datos, la automatización de procesos o la atención al cliente. Capaces de producir texto, imágenes, código o incluso audio y vídeo a partir de instrucciones en lenguaje natural las IAs generativas tienen, sin lugar a dudas, un papel destacado en esta evolución.

El pistoletazo de salida lo dio Chat GPT que, desarrollada por OpenAI, superó los cien millones de usuarios activos mensuales apenas dos meses después de su lanzamiento. Pero no es la única. Microsoft Copilot; Google Gemini, integrada en productos como Workspace y Android; Claude, de Anthropic, con un enfoque

centrado en la seguridad y la transparencia; y LLaMA, de Meta, que apuesta por un modelo de desarrollo más abierto, son otras opciones destacadas, a las que se suman iniciativas de empresas como Mistral AI, Cohere, Stability AI y otras muchas que nutren un ecosistema en expansión constante.

Sin embargo, este despliegue masivo de la IA generativa ha multiplicado también los riesgos asociados a su uso. Se han documentado casos de envenenamiento de datos que afectan a la calidad y fiabilidad de los modelos; fugas de información a través de las respuestas generadas; manipulación de las instrucciones de entrada para inducir comportamientos indebidos; y explotación maliciosa por parte de actores que emplean estas herramientas con fines como la ingeniería social, la creación de malware o la desinformación.

Este escenario deja clara la necesidad de proteger tanto los modelos de IA como toda su cadena de suministro, desde los datos con los que se entrenan hasta los entornos donde se ejecutan. La seguridad de la inteligencia artificial se ha convertido en un nuevo frente de batalla que las organizaciones deben integrar en su estrategia de ciberseguridad.

> La adopción masiva de modelos generativos ha multiplicado los riesgos asociados a la inteligencia artificial.

Principales vectores de ataque contra la IA

Ya hemos comentado que los vectores de ataque contra la IA se multiplican. Entre los más relevantes se encuentran los ataques adversariales, capaces de engañar al sistema mediante alteraciones casi imperceptibles en las entradas. Un ejemplo real está relacionado con los vehículos autónomos: se ha demostrado que basta con modificar ligeramente una señal de tráfico para que el sistema de visión interprete una señal de “stop” como una de “límite de velocidad”, generando un riesgo evidente.

Otro frente delicado es el del *model stealing* y el *model inversion*. En el primer caso, los atacantes replican modelos propietarios mediante consultas repetidas a sus APIs públicas, como ocurrió en 2020 con servicios comerciales de Google o Amazon. En el segundo, se logra reconstruir información sensible del proceso de entrenamiento: un estudio académico demostró cómo era posible recrear rostros de pacientes a partir de un modelo médico, lo que plantea serias implicaciones para la privacidad.

El *data poisoning* se ha convertido también en una técnica eficaz para degradar el comportamiento de los modelos, especialmente si estos se entrenan de forma continua con datos de origen no controlado. El ejemplo más recordado es el del chatbot Tay, de Microsoft, que en

Datos para dimensionar el reto
Check Point Research

Según el informe State of AI Cyber Security elaborado por Check Point Research:

- **Más del 51 % de las redes corporativas usan herramientas de IA cada mes.**
- **El 1,25 % de los prompts enviados desde dispositivos empresariales presenta alto riesgo de fuga de datos, y otro 7,5 % contiene información potencialmente sensible.**
- **Los atacantes están desarrollando modelos LLM maliciosos como WormGPT, FraudGPT o DarkGPT, y servicios como GoMailPro automatizan campañas de phishing por 500 dólares al mes.**
- **Se han registrado fraudes que emplean deepfakes en tiempo real para suplantar identidades por vídeo o voz, incluyendo un caso con pérdidas de 20 millones de libras esterlinas para una empresa británica.**

pocas horas adoptó un discurso ofensivo tras ser manipulado por usuarios maliciosos.

Más recientemente, ha ganado protagonismo la llamada *prompt injection*, una técnica que consiste en introducir instrucciones ocultas dentro de un mensaje aparentemente inocuo para alterar el comportamiento del modelo. Se han registrado casos en los que este tipo de manipulación ha permitido eludir restricciones o acceder a información sensible, lo que supone una amenaza para cualquier sistema que integre modelos generativos en interfaces o flujos automatizados.

Por último, no se puede obviar la creciente preocupación en torno a las vulnerabilidades presentes en modelos de código abierto. La colaboración abierta ha impulsado la innovación, pero también ha facilitado la aparición de nuevos riesgos. En 2023, la plataforma Hugging Face tuvo que retirar modelos subidos por usuarios que contenían código malicioso oculto diseñado para ejecutarse en el momento de la carga. Este incidente obligó a reforzar las medidas de control y supervisión en repositorios públicos.

Enfoques y soluciones tecnológicas para proteger la IA

El despliegue seguro de sistemas de inteligencia artificial exige la incorporación de nuevas capas de protección. Una de las estrategias es la monitorización del comportamiento del modelo en tiempo real, también conocida como *ML observability*, una práctica que permite detectar desviaciones, anomalías o manipulaciones a partir de la evolución de las predicciones, la deriva en los datos o cambios inesperados en el rendimiento del sistema.

Junto a esta monitorización, han cobrado relevancia las técnicas de evaluación de robustez, que someten a los modelos a pruebas frente a ataques adversariales simulados. Algunas herramientas incluso escanean las entradas en tiem-

po real para identificar patrones sospechosos antes de que sean procesados por el modelo.

En el ámbito de la privacidad, están ganando terreno el cifrado homomórfico y la computación confidencial como vías para proteger los datos durante la inferencia. Por su parte, el aprendizaje federado permite entrenar modelos sin necesidad de centralizar la información, lo que reduce significativamente el riesgo de exposición.

Otro pilar clave para proteger la IA es el control de acceso y la gestión de identidades, especialmente en entornos donde los modelos son accesibles a través de APIs o interfaces compartidas. Aplicar el principio de Zero Trust, acompañado de políticas de gestión de privilegios y auditoría del uso, se vuelve esencial para evitar abusos o accesos indebidos.

Finalmente, los mecanismos de trazabilidad están adquiriendo un peso creciente. La capacidad de registrar todas las interacciones con un modelo, mediante *audit logs* o técnicas de *watermarking*, no solo permite detectar alteraciones o usos no autorizados, sino que también facilita el cumplimiento normativo y la atribución de responsabilidades en caso de incidente.

> Proteger los modelos de IA generativa, sus datos y su cadena de suministro es ya una prioridad crítica para la ciberseguridad empresarial.

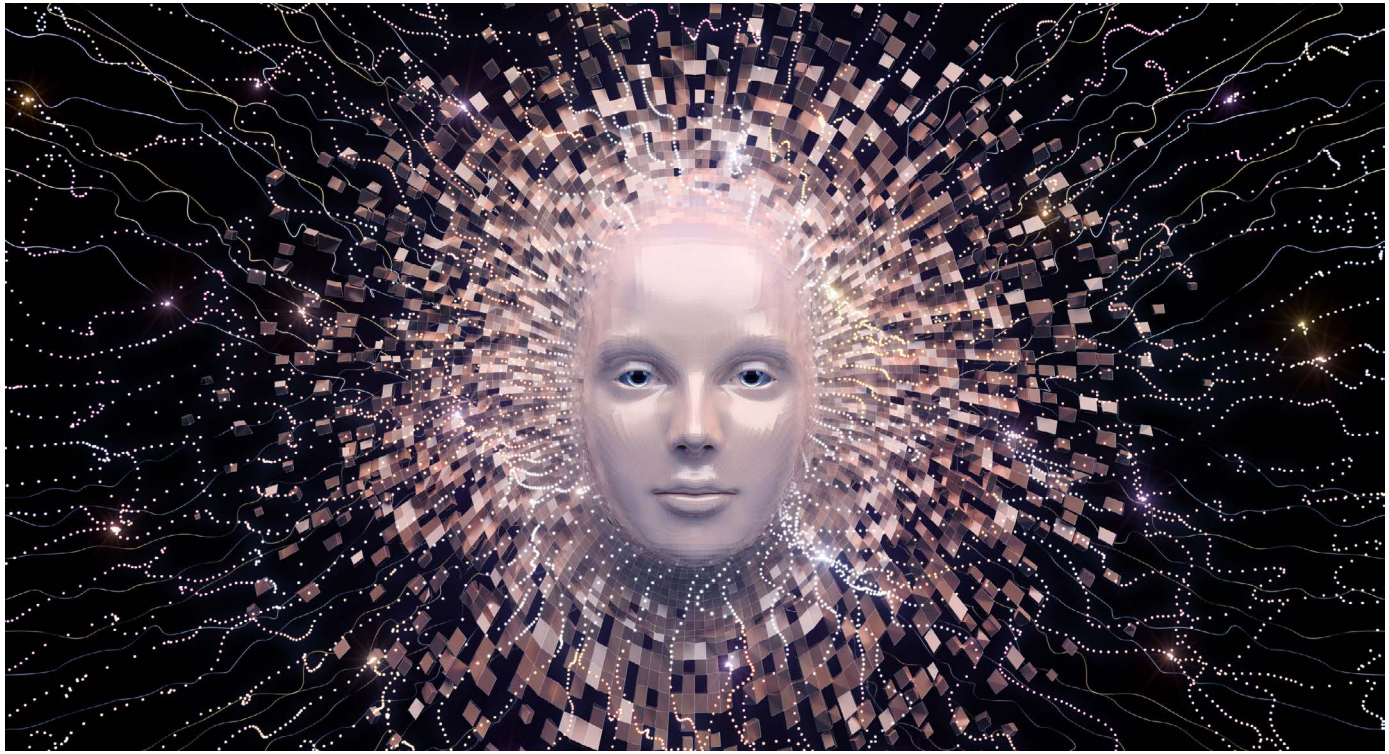
Agentes IA. Nos son humanos, pero actúan como tales

Una de las grandes novedades que está marcando el avance de la inteligencia artificial en 2025 es la irrupción de los llamados agentes de IA. A diferencia de los modelos generativos tradicionales, que responden a una entrada puntual, estos agentes están diseñados para operar de forma autónoma, persistente y proactiva en entornos digitales. Pueden descomponer objetivos complejos en tareas, tomar decisiones encadenadas y ejecutarlas sin intervención humana directa. Esta evolución no sólo representa un salto en eficiencia operativa, sino que introduce un nuevo tipo de entidad digital en el ecosistema tecnológico, con implicaciones profundas tanto para la productividad como para la seguridad.

Desde una perspectiva de ciberseguridad, la presencia de agentes autónomos plantea retos inéditos. Al tener capacidad para interactuar con múltiples sistemas, acceder a datos sensibles o incluso modificar configuraciones, estos agentes deben ser tratados como actores privilegiados, aunque no sean humanos. Esto exige nuevas estrategias para su identificación, autenticación, autorización y monitorización dentro de las arquitecturas corporativas. Su dinamismo y autonomía dificultan aplicar controles tradicionales: pueden cambiar de contexto, aprender comportamientos o establecer conexiones inesperadas. Por ello, los expertos reclaman marcos de gobernanza específicos para su desarrollo y despliegue, así como mecanismos de auditoría continua

que permitan supervisar sus decisiones y detectar desviaciones o abusos.

Además, si estos agentes son maliciosos o comprometidos —por ejemplo, a través de un modelo envenenado o una instrucción manipulada—, el daño potencial es enorme. La combinación de autonomía, persistencia y acceso puede facilitar desde movimientos laterales hasta filtraciones de datos o sabotaje de operaciones críticas. No es casualidad que varios analistas ya los consideren una futura herramienta de ataque tan peligrosa como el malware tradicional.



Fabricantes y herramientas destacadas

La RSA Conference 2025, uno de los eventos de ciberseguridad más importantes del mundo, ha confirmado que la seguridad de la inteligencia artificial se ha convertido en una de las prioridades estratégicas del sector. Un número creciente de fabricantes ha centrado sus lanzamientos en soluciones específicas para proteger modelos, agentes, datos y procesos impulsados por IA.

Aunque su propuesta más ambiciosa, Cisco AI Defense, no se presentó en el marco de la RSA sino meses antes, en enero, esta solución es un claro ejemplo de cómo estamos asistiendo a una oleada de propuestas que combinan tecnologías avanzadas con enfoques de seguridad desde el diseño. AI Defense se plantea como una plataforma integral de protección en entornos multimodelo, multinube y multiagente, y responde a un problema clave en el despliegue de IA a gran escala: la dispersión de estándares de seguridad entre proveedores. AI Defense propone una capa unificada de protección, aplicable a modelos independientemente de su origen, que permite establecer políticas de seguridad coherentes, prevenir la exposición de

La RSA Conference 2025 ha marcado un punto de inflexión: la seguridad de la inteligencia artificial ha pasado de ser una preocupación emergente a convertirse en eje central de las estrategias de fabricantes y proveedores.

datos sensibles y facilitar la validación rápida de los modelos, pasando de semanas a segundos. Durante la RSA Conference, varias compañías presentaron soluciones que refuerzan la idea de que proteger la IA y protegerse con IA son ya ejes prioritarios para la industria. Microsoft, por ejemplo, anunciaba una mejora en la seguridad de sus plataformas de IA con Azure AI Security, que incorpora protección del ciclo de vida del modelo, control de acceso, auditoría y detección de entradas maliciosas.

En el entorno de Google Cloud, la plataforma Vertex AI ofrece mecanismos de explicabilidad, gobernanza segura y detección automatizada de desviaciones, todo ello integrado con herramientas como AI Explanations o AI Governance Toolkit.

IBM apuesta por watsonx.governance, una suite orientada a la gobernanza responsable que automatiza la documentación de los modelos, audita inferencias, detecta sesgos y valida el cumplimiento normativo, con integración en flujos de MLOps.

OpenAI, por su parte, ha reforzado su API empresarial y la versión corporativa de ChatGPT con cifrado de extremo a extremo, autenticación avanzada y controles contra *prompt injection*, fugas de contexto y otros usos maliciosos.

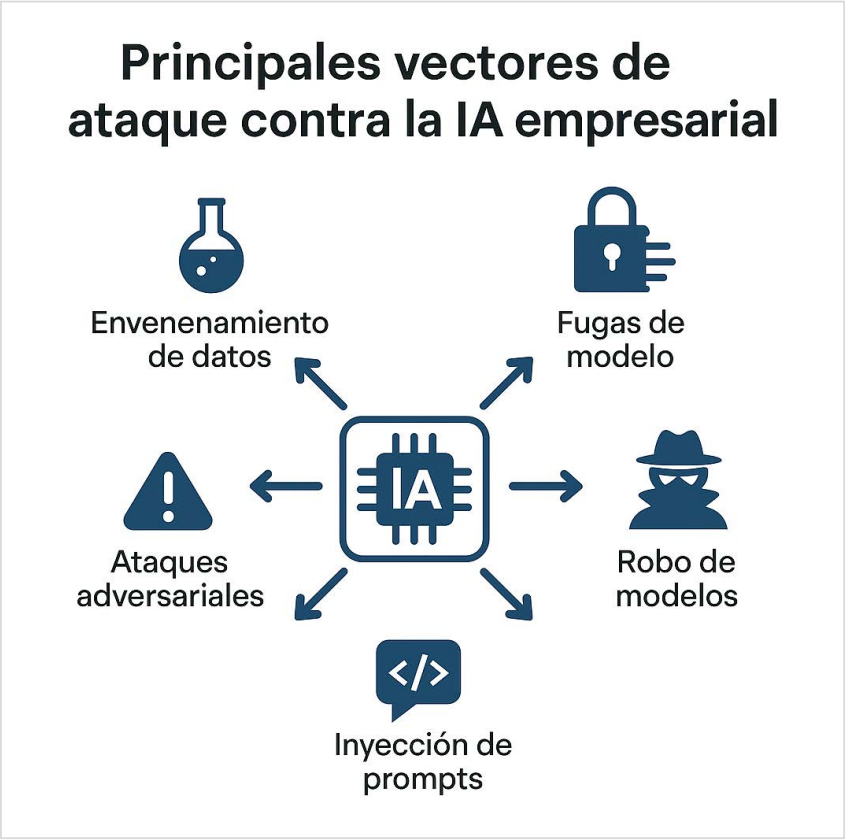
Numerosas startups han aprovechado la RSA para presentar soluciones innovadoras en este terreno. HiddenLayer, Robust Intelligence, CalypsoAI o Lakera han destacado por abordar con precisión riesgos específicos como los ataques adversariales, el *model stealing*, la inyección de prompts o la validación de modelos antes de su puesta en producción. Lakera, por ejemplo, presentó nuevos mecanismos defensivos para proteger agentes LLM integrados en entornos empresariales.

En paralelo, varias compañías han apostado por soluciones basadas en agentes autónomos impulsados por IA con funciones de ciberseguridad. Abnormal AI ha lanzado dos agentes diseñados para entrenamiento en concienciación y análisis de datos de seguridad, mientras que ArmorCode presentó a Anya, una IA conversacional para AppSec capaz de correlacionar datos desde más de 285 integraciones. Arctic Wolf, en colaboración con Anthropic, dio a conocer Cipher, un asistente de seguridad alimentado por LLMs para generar análisis expertos

en tiempo real. Este tipo de soluciones demuestra cómo la IA no solo genera nuevos riesgos, sino también herramientas capaces de fortalecer la defensa digital.

En el ámbito de la infraestructura crítica, Cisco ha destacado con nuevas capacidades para gestionar el riesgo en la cadena de suministro de IA y proteger aplicaciones que utilizan modelos generativos, incluyendo la creación de Foundation AI, un equipo dedicado a acelerar una adopción segura de esta tecnología. NVIDIA, por su parte, ha anunciado que llevará la ciberseguridad en tiempo de ejecución a todas las "fábricas de IA" con su nueva arquitectura DOCA Argus.

Desde una óptica centrada en la criptografía, Utimaco ha ampliado su gama de protección con Quantum Protect y HSMs preparados para entornos con IA, integrándose ahora también con IBM Cloud en el sector financiero. Su alianza con Bloombase refuerza la protección de datos en reposo mediante cifrado de alto rendimiento y control de acceso basado en hardware.



La seguridad de agentes autónomos también ha sido un foco destacado. EQTY Lab presentó AI Guardian, un sistema de cumplimiento en tiempo real desarrollado junto a Intel y NVIDIA para garantizar que los agentes actúan conforme a políticas y estándares de seguridad. Sentra, por su parte, ha lanzado una solución para controlar cómo los agentes acceden y manejan datos sensibles en flujos de trabajo basados en IA.

En el plano operativo, Securonix y Vectra AI han incorporado agentes generativos capaces de actuar como analistas virtuales, con el objetivo de aliviar la carga de los equipos SOC y permitir respuestas más rápidas y precisas ante incidentes.

Por último, herramientas como Qualys TotalAI y Cyberhaven han extendido la protección a toda la cadena MLOps, con escaneos en fase de desarrollo, análisis del flujo de datos generados y cobertura ante amenazas multimodales, evidenciando un enfoque holístico que va más allá del perímetro técnico de los modelos.

Normativas y marcos de referencia emergentes

La rápida adopción de la IA ha acelerado también la creación de marcos normativos y guías

internacionales que buscan garantizar su uso ético, seguro y transparente. Aunque aún en fases distintas de aplicación, estas iniciativas ofrecen una hoja de ruta clara para las organizaciones.

El Reglamento Europeo de Inteligencia Artificial (IA Act) es la primera legislación integral sobre IA a nivel mundial. Clasifica los sistemas según su nivel de riesgo y establece requisitos específicos para los considerados de alto riesgo, como los usados en sanidad, infraestructuras críticas o procesos de selección de personal. Además, incluye obligaciones sobre transparencia, supervisión humana, ciberseguridad y calidad de los datos.

En Estados Unidos, el Instituto Nacional de Estándares y Tecnología (NIST) ha publicado un marco voluntario para la gestión de riesgos en IA. Basado en cuatro funciones clave —gobernar, mapear, medir y gestionar— el NIST AI RMF promueve prácticas centradas en la seguridad, la fiabilidad, la mitigación de sesgos y la explicabilidad de los modelos.

La Agencia Europea de Ciberseguridad (ENISA) también ha publicado múltiples informes centrados en los desafíos de seguridad en IA, con análisis de amenazas, buenas prácticas y recomendaciones para el despliegue seguro de modelos. ENISA colabora además con los Estados miembro en el diseño de estrategias nacionales que integren IA y ciberseguridad.

Por último, el G7 ha impulsado el Hiroshima AI Process, una iniciativa que, aunque no vinculante, pretende armonizar estándares entre democracias industriales en torno a principios como la transparencia, la seguridad, el respeto a los derechos humanos y la supervisión responsable del uso de la inteligencia artificial.

Datos para dimensionar el reto
Orca Security

Según el informe 2024 State of AI Security Report elaborado por Orca Security:

- **El 56 % de las organizaciones ya han adoptado modelos de IA para aplicaciones personalizadas**
- **El 62 % han desplegado al menos un paquete de IA con una vulnerabilidad conocida (CVE), aunque en su mayoría son de riesgo bajo o medio (CVSS medio de 6,9)**
- **El 20 % de las organizaciones que usan OpenAI tienen claves de acceso expuestas en ubicaciones no seguras**
- **El 81 % de las claves de OpenAI expuestas estaban almacenadas en el historial de commits de repositorios públicos**
- **El 98-100 % de las organizaciones no han configurado cifrado en reposo con claves autogestionadas en servicios como Azure OpenAI, Vertex AI o Amazon SageMaker**