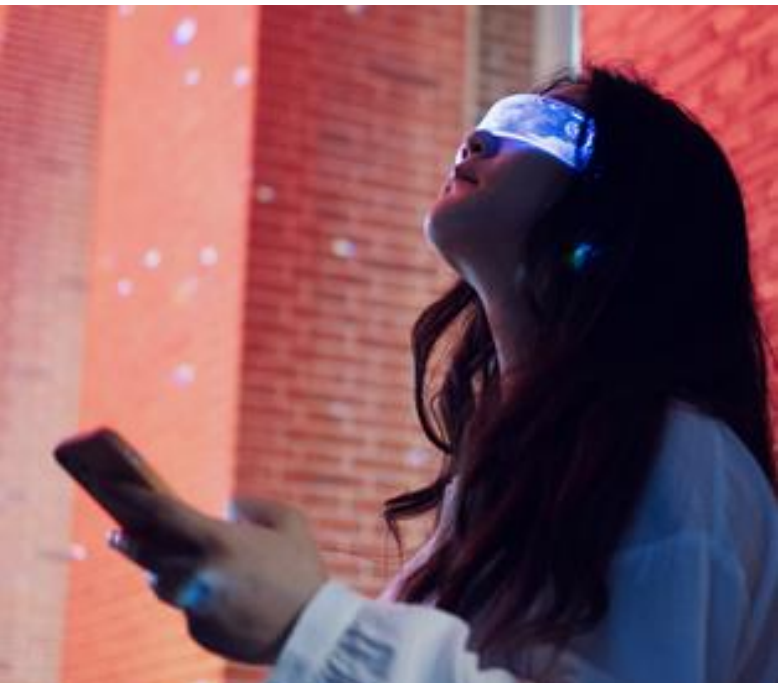
Agentes de cambio: El auge de la IA autónoma en ciberseguridad

Se abre un nuevo capítulo en la ciberseguridad. Uno en el que las máquinas siguen órdenes humanas, pero trazan su propio rumbo. Este cambio está impulsado por la IA con agentes: sistemas autónomos capaces de razonar, planificar y actuar de forma independiente para alcanzar objetivos complejos. Los agentes de IA permiten a las empresas repensar y reimaginar su forma de trabajar.

En el contexto de la ciberdefensa, estos agentes no solo son herramientas más eficientes, sino también colaboradores emergentes. Imagine sistemas inteligentes que detectan amenazas en tiempo real, coordinan respuestas en todas las redes, investigan vulnerabilidades y ajustan sus tácticas a medida que cambian las condiciones. No esperan instrucciones. Actúan.

Esto es más que una mejora en la capacidad. Es un cambio en la forma en que se diseña y gestiona la ciberseguridad. Y plantea importantes preguntas:

- ¿Cómo gestionamos lo que no programamos explícitamente?
- ¿Cómo se ve la responsabilidad cuando el actor es un algoritmo?
- Y, lo más importante, ¿cómo lideramos en una era donde agentes inteligentes operarán junto a nosotros, y a veces, por delante de nosotros?



Esta serie del Instituto de Innovación en Ciberseguridad y Riesgos de PwC explora la nueva frontera de la IA agéntica en ciberseguridad: las oportunidades, las amenazas y el liderazgo necesarios para dar forma al futuro. Aquí un vistazo a lo que nos espera:

1. **IA agéntica: la próxima frontera de la ciberdefensa**
   Los CISO y líderes de seguridad de hoy deberían repensar los roles, las responsabilidades y los riesgos en una era en la que los compañeros de equipo digitales toman decisiones de misión crítica.

2. **Agentes de IA: ¿Tu próxima amenaza interna?**
   Los sistemas autónomos con amplio acceso y poder de toma de decisiones plantean un nuevo tipo de riesgo interno, especialmente si se ven comprometidos o desalineados.

3. **Quién tiene el control: gestión de los riesgos cibernéticos de los agentes de IA**
   Desde las restricciones y el monitoreo hasta los apagados elegantes, ¿cómo mantenemos a la IA agente responsable?

4. **Guardianes cibernéticos: IA agéntica como el equipo azul de próxima generación**
   Conoce a los agentes diseñados para la defensa: cómo se utiliza la IA autónoma para monitorear, responder y burlar a los atacantes en tiempo real.

5. **Cuando los equipos de IA pierden el rumbo: cómo gestionar la amenaza interna**
   Los agentes de IA colaborativos pueden ayudar a resolver problemas complejos, pero el comportamiento emergente o las fallas en la coordinación podrían generar problemas.

6. **Shadow Ops: ciberamenazas ofensivas impulsadas por IA agéntica**
   Cómo los atacantes podrían utilizar la IA agéntica en operaciones futuras, desde el reconocimiento inteligente hasta la entrega dinámica de carga útil.

La IA agéntica no solo está cambiando las herramientas que usamos; está reescribiendo las reglas de juego en ciberseguridad. Nos adentramos en un mundo donde los sistemas autónomos pueden defender, atacar, adaptarse y evolucionar más rápido que los humanos. Para quienes defienden, esto va más allá de mantenerse al día. Se trata de reinventar lo posible.

Las organizaciones que prosperen en esta nueva era no solo adoptarán la IA. Desarrollarán estrategias, culturas y equipos digitales que piensen y actúen junto a nosotros como compañeros. **Esta serie es tu mapa hacia ese futuro. La era de los agentes ya ha comenzado. Liderémosla.**

Parte 1

IA agéntica: la próxima frontera de la ciberdefensa

Cómo los CISO pueden preparar a su personal

Los programas de ciberseguridad actuales se basan en un modelo operativo centrado en el ser humano, donde los analistas inician acciones, definen prioridades y responden manualmente a las amenazas. Este modelo ha sido muy útil para las empresas, especialmente gracias a que la IA ha ayudado a avanzar y automatizar el análisis de amenazas. Sin embargo, ya no es suficiente por sí solo para responder a ataques más sofisticados y específicos.

El auge de la IA agentica —sistemas autónomos y orientados a objetivos, capaces de tomar decisiones y ejecutar tareas con mínima intervención humana— marca un avance transformador para la ciberdefensa. Estos sistemas de IA ya no son solo herramientas que ayudan con el análisis; se están convirtiendo en compañeros de equipo digitales que pueden actuar de forma independiente, colaborar con equipos humanos e incluso iniciar respuestas de seguridad.

Para los CISO, esto no es solo una evolución técnica. Es un cambio cultural y organizativo que exige una nueva forma de pensar para ganarse la confianza y la aceptación de los equipos de seguridad.

Forjando una alianza en ciberdefensa

Con la IA con agentes, los CISO están entrando en una nueva era de equipos híbridos donde los agentes trabajan activamente junto con analistas humanos. Los agentes pueden detectar amenazas, clasificar incidentes, orquestar respuestas y aprender continuamente de sus entornos.

El rol humano ahora cambia de **ejecutor a supervisor, estratega y responsable ético.**

Esta transición desafía las suposiciones arraigadas sobre roles y responsabilidad. Los profesionales de la ciberseguridad deben desarrollar un nuevo tipo de alfabetización digital, pasando del uso de herramientas a la gestión de sistemas autónomos que operan de forma independiente.

Cómo los CISO pueden movilizar a sus equipos

Los equipos de seguridad deben aprender a colaborar y comunicarse con ellos —a menudo en su lenguaje natural— y comprender sus límites de decisión. Además, deben establecer un proceso de gobernanza para supervisar e identificar las medidas a tomar cuando se rompe la confianza.

Al considerar la incorporación de un enfoque de IA agente en su programa cibernético, los CISO pueden preparar a sus equipos de seguridad mediante:

Imagina sistemas inteligentes que detectan amenazas en tiempo real, coordinan respuestas en todas las redes, investigan vulnerabilidades y ajustan sus tácticas a medida que cambian las condiciones.

Centrándose en la colaboración humanos-IA, con los humanos al mando.

Capacita a tu fuerza laboral cibernética y desmitifica la IA agéntica. Comunica claramente las capacidades y limitaciones de la IA agéntica, y cómo puede colaborar con los humanos para mejorar la ciberseguridad. Que los colaboradores comprendan el propósito y los beneficios de la IA agéntica puede ayudar a alinear a los equipos con tu estrategia de ciberseguridad y objetivos de resiliencia.

Capacita a sus equipos para desarrollar casos de uso de IA con agentes. Establece un entorno donde los colaboradores puedan idear y presentar casos de uso para su desarrollo en el ámbito de la inteligencia artificial y así resolver problemas que les preocupan. Fomenta una cultura que promueva el intercambio de conocimientos y la resolución de problemas en torno al uso de herramientas cibernéticas de IA con agentes para resolver desafíos reales de ciberseguridad.

Hacer que la IA agéntica sea responsable y confiable en cada paso.

Establece controles robustos de gestión de identidad y acceso de IA. Los sistemas de IA agéntica necesitarán acceso y, en ocasiones, incluso "identidades" para establecer operaciones de ciberseguridad responsables. Otorga a la IA agéntica solo los privilegios necesarios para sus tareas, siguiendo los procesos y controles estándar de identidad y acceso (IAM). Supervisa constantemente su comportamiento y realiza revisiones periódicas de acceso e identidad.

Integra los principios de IA responsable desde el principio. La clave para la adopción y el éxito de la IA es generar confianza. Al aplicar los principios de la IA Responsable, puedes ayudar a generar confianza entre líderes, colaboradores, clientes y otras partes interesadas. Además, es importante supervisar continuamente las regulaciones y estándares del sector, especialmente dada la rápida evolución de la IA.

Anticipando cambios de paradigmas y roles.

De una defensa impulsada por analistas a una defensa reforzada por agentes. Los analistas humanos ascenderán a roles cada vez más estratégicos, como la supervisión, el modelado de amenazas y la estrategia de riesgos.

De roles fijos a una colaboración fluida. Los equipos evolucionarán dinámicamente, y los agentes asumirán diferentes responsabilidades según las necesidades.

De herramientas aisladas a ecosistemas interoperables. Los agentes de IA prosperarán en entornos donde se puedan integrar herramientas, datos y plataformas, lo que impulsará a los líderes cibernéticos a priorizar la interoperabilidad y la preparación para la IA en su conjunto tecnológico.

La IA agéntica no está reemplazando al profesional de la ciberseguridad, sino que está redefiniendo lo que significa serlo. Los CISO que adopten este cambio ahora podrán posicionar a sus organizaciones a la vanguardia de la defensa aumentada por IA. El futuro pertenecerá a quienes vean a estos agentes no como amenazas, sino como compañeros de equipo.

Parte 2

Agentes de IA: ¿Tu próxima amenaza interna?

El entusiasmo en torno a los agentes de IA es palpable. Según la encuesta de PwC sobre agentes de IA, el 79% de los altos ejecutivos afirma que sus empresas ya los están adoptando. Las organizaciones se apresuran a implementar compañeros de equipo digitales que puedan automatizar tareas rutinarias como la reserva de reuniones y el procesamiento de facturas. A medida que esta tecnología evoluciona, las capacidades se amplían para admitir funciones más complejas, como la orquestación de flujos de trabajo completos.

Sin embargo, incluso mientras los ejecutivos buscan aprovechar al máximo el valor de los agentes, la ciberseguridad es el principal desafío al que se enfrentan. Entre los numerosos riesgos emergentes, los agentes de IA introducen una nueva forma distintiva de amenaza interna. Pueden operar con acceso a sistemas y datos confidenciales, sin la supervisión que se suele aplicar a los usuarios humanos.

El término "amenaza interna" se refiere tradicionalmente al riesgo que representan las personas dentro de una organización, como empleados, contratistas o socios comerciales, que utilizan su acceso a los sistemas, datos o recursos de la organización para causar daños, intencional o involuntariamente. Históricamente, estos agentes internos han causado daños a través de una amplia gama de acciones, desde fraude y robo hasta sabotaje y espionaje. En ocasiones, han sido influenciados por actores amenazantes de estados-nación u otros adversarios.

79%

de los altos ejecutivos afirman que sus empresas ya están adoptando agentes.

Fuente: Encuesta de Agentes de IA de PwC, mayo de 2025.

How agents can turn into insider threats

En teoría, los agentes de IA con acceso comparable podrían participar o ser manipulados para realizar las mismas actividades que los humanos. En el momento en que las organizaciones otorgan a los agentes de IA capacidades similares a las humanas, básicamente crean nuevos empleados con acceso al sistema y poder de decisión.

A diferencia del *software* tradicional que sigue reglas predefinidas, los agentes interpretan instrucciones, toman decisiones y ejecutan acciones de forma autónoma en tiempo real. Este nivel de independencia es precisamente lo que los convierte en un vector de amenaza potencial. Pueden realizar sus actividades a gran escala y mucho más rápido que los humanos, lo que aumenta el riesgo que representan. Esto hace que la supervisión humana y el establecimiento de barreras de seguridad sean aún más esenciales.

Considera este escenario: Utilizas un agente de IA para resumir los correos electrónicos no leídos y sugerir información crítica que requiera tu atención. Un atacante inyecta un mensaje malicioso para alterar los objetivos e instrucciones de tu agente de IA, lo que provoca que envíe la información crítica al correo electrónico del atacante.

El agente no es malicioso; ha sido obligado a realizar acciones maliciosas. Al agregar agentes de IA, tienes un nuevo vector de amenaza interna en tu red. A diferencia de los humanos, no se expresan abiertamente y no presentan indicios de manipulación.

Actualmente, existe una brecha en la gestión de este tipo de amenazas, ya que aún no contamos con un marco holístico claro y completo para proteger a los agentes de IA. Si bien contamos con décadas de experiencia con amenazas internas humanas y ciberseguridad tradicional, la seguridad de la IA con agentes es un territorio prácticamente inexplorado. La realidad es que estamos comenzando un experimento masivo en producción.

¿Qué podemos hacer hoy?

Antes de entregar más llaves a los agentes de IA, las organizaciones necesitan:

Tratar a los agentes de IA como a los demás usuarios
Implementar un sistema de mínima intervención, monitoreo de actividad y auditorías periódicas de los niveles de acceso y las necesidades del negocio.

Establecer límites claros
Definir lo que los agentes pueden y no pueden hacer, con límites estrictos para acciones de alto riesgo. Aplicar límites de velocidad y bucle para evitar agentes descontrolados.

Implementar controles con intervención humana
Añadir controles y contrapesos a decisiones o transacciones financieras importantes.

Monitorear el comportamiento del agente
Establecer revisiones independientes fuera de las operaciones del agente, como un agente observador, para ayudar a detectar anomalías y patrones inesperados. Aplicar continuamente la puntuación de riesgos para anticiparse a las posibles amenazas.

Imponer un modelo de código ético para agentes
Capacitar a los agentes no solo en sus tareas, sino también en los valores y la misión de la empresa, además de las políticas y normas de negocio de la organización.

Planificar el inventario de agentes y su revisión periódica
Programe revisiones periódicas para reevaluar el acceso y los privilegios de cada agente, especialmente a medida que sus roles y capacidades evolucionan.

Incorporar la paranoia como característica
Programar agentes con mayor conciencia para que detecten indicaciones sospechosas. Deberían preguntarse: "¿Esa instrucción fue normal o podría ser un truco?". Esta mayor conciencia podría ser especialmente útil en entornos con mayor riesgo de fraude o ingeniería social.

Implementar tokens canarios
Agregar objetos canarios en entornos con los que los agentes están entrenados para no interactuar debido a su sensibilidad. Activar alertas cuando los agentes accedan a estos objetos, lo que indica una posible manipulación.

El potencial de los agentes de IA es inmenso. Para que las organizaciones aprovechen al máximo su valor, es fundamental protegerlos. Como ocurre con cualquier ciberamenaza, la pregunta es cuándo, y no si, los agentes de IA serán atacados. Ser proactivo con las capacidades de monitoreo, aplicar los controles necesarios y mantener prácticas responsables de IA son clave para mitigar estos riesgos y facilitar el despliegue de agentes.

Parte 3

Quién tiene el control: gestión de los riesgos cibernéticos de los agentes de IA

La IA con agentes ya está aquí, y las empresas ya están poniendo a los agentes a trabajar. Según la Encuesta de Agentes de IA, más de la mitad de quienes la han adoptado esperan ampliar su implementación en un plazo de seis meses, centrándose en la atención al cliente y el soporte (57%), las ventas y el marketing (54%) y las TI y la ciberseguridad (53%)[1].

Estos agentes aprenden de la experiencia, recuerdan interacciones pasadas e incluso se coordinan con otros agentes. Equipados con API internas, complementos e integraciones directas con el sistema, pueden actualizar registros, aprobar flujos de trabajo y activar procesos posteriores.

A medida que los agentes pasan de proyectos piloto a ser trabajadores habituales, amplían rápidamente la superficie de ataque empresarial. Las prácticas de IA responsable pueden ayudar a las organizaciones a construir sistemas éticos y transparentes. Y las prácticas de IA segura transforman la seguridad de un parche adicional a una obligación de diseño. Aplicadas a la IA agente, estas prácticas pueden guiar la concesión de autonomía y fortalecer la seguridad desde cero. Pero, al igual que otras integraciones de plataformas y aplicaciones, cuantos más permisos, datos y herramientas posean, mayor será el riesgo de comportamientos imprevistos y efectos en cascada. Los agentes pueden ampliar la superficie de ataque y crear cadenas de problemas de seguridad — desde permisos de acceso e identidades privilegiadas hasta protección de datos— que desafían las estrategias tradicionales de ciberseguridad y gobernanza de la IA. Además, están surgiendo nuevos tipos de explotación.

Ejemplo: Un agente con acceso al correo electrónico y al calendario de un usuario para la programación automatizada es manipulado para reenviar información confidencial por correo electrónico a través de la nota de una confirmación de reunión imaginaria.Simplemente ampliar los controles de seguridad actuales no será suficiente. Gestionar la IA agentica exige protecciones específicas adaptadas a las arquitecturas y los vectores de amenaza que introducen estos sistemas.

¿Cómo pueden los CISO abordar los nuevos riesgos que plantean los agentes?

A diferencia de los humanos, los agentes de IA carecen de criterio y comprensión de las consecuencias y la responsabilidad. Sin embargo, se les puede confiar la toma de decisiones que pueden tener consecuencias de gran alcance a una velocidad que ningún humano podría replicar. Los ejecutivos aún se muestran recelosos de confiar en que los agentes actúen de forma autónoma en las interacciones con los clientes (25%) y con los empleados (22%).

Los responsables de seguridad se enfrentan a un desafío inmediato. No existe un marco de seguridad establecido para agentes de IA (hasta la fecha) que oriente sobre cómo supervisar y controlar los distintos comportamientos de estos agentes.

Marcos como CSA MAESTRO y OWASP Top Ten ofrecen orientación sobre la mitigación de amenazas de agentes, y otros están ganando terreno. En cualquier caso, esto no debería impedirle adoptar proactivamente las mejores prácticas del sector para comenzar a realizar evaluaciones de riesgos y actualizarlas.

Comprender la estrategia de tu organización para la adopción de IA agéntica

Como líder de seguridad, debe colaborar con los líderes empresariales y su director de sistemas de información (CIO) para determinar el plan de su organización para la adopción de IA con agentes. Esto incluye una revisión de las plataformas y herramientas preferidas, planes para impulsar el desarrollo interno, facilitar el desarrollo ciudadano o contratar agentes externos.

Obtener claridad estratégica es esencial para evaluar las implicaciones de seguridad, gobernanza y gestión (y orquestación) en cada nivel. Con esta perspectiva, su equipo directivo puede anticipar mejor los riesgos potenciales, definir controles de seguridad adecuados e integrar la seguridad en el ciclo de vida de la IA desde el principio.

[1]Las siguientes estadísticas, a las que se hace referencia en este documento, provienen de esta encuesta.

Identificar las amenazas que enfrentarás

La adopción de IA agéntica amplía y modifica significativamente el panorama de amenazas y los riesgos cibernéticos asociados que tu empresa afrontará. Comienza a analizar las amenazas desde perspectivas históricas y novedosas:

Nuevas amenazas que plantean los agentes de IA: los agentes de IA introducen amenazas únicas que implican comportamientos tortuosos (p. ej., colusión de múltiples agentes cuando varios agentes se coluden y eluden defensas como los controles de acceso) y escenarios de amenazas que implican la explotación de agentes de IA para causar riesgos cibernéticos (p. ej., corrupción de la memoria del agente que implica manipulación no autorizada, envenenamiento o degradación accidental del almacén de memoria de un agente de IA, lo que lleva a violaciones de la política de seguridad).

Amenazas relevantes para GenAI: Las capacidades GenAI de la IA agéntica pueden explotarse para exponer diversos riesgos cibernéticos. Esto incluye escenarios de amenaza como ataques de inyección de indicaciones, en los que un atacante manipula la indicación de entrada para alterar el comportamiento del modelo, eludir sus instrucciones o provocar acciones no deseadas.

Ciberamenazas tradicionales: La IA agéntica puede considerarse un activo de información común que puede ser blanco de amenazas maliciosas mediante *exploits* de vulnerabilidades, ataques DDoS, *exploits* de configuración incorrecta o *exploits* de API. Además, puede ser objeto de ataques de *malware* o ingeniería social, así como de manipulación interna maliciosa.

Adaptar tus controles para mitigar los riesgos de la IA agente

Al alinear la seguridad de GenAI con la IA Responsable, prepárate para evaluar los riesgos de la IA agéntica y fortalecer los controles para mantener la resiliencia. Comienza evaluando la idoneidad de los controles existentes para gestionar las amenazas y los riesgos aplicables, a la vez que evalúas dónde se necesitan nuevas capacidades y supervisión humana para reforzar las ciberdefensas.

- **Establecer espacios seguros para la adopción.** Permitir el uso de un número razonable y manejable de ecosistemas de IA empresarial con agentes dentro de la organización (p. ej., el sistema operativo de agentes de PwC, Google Agentspace, Microsoft Copilot Studio, Salesforce Agentforce) para obtener visibilidad de los agentes y gestionarlos de forma segura. Los profesionales de TI llevan años lidiando con la TI en la sombra. Establecer vías de IA aprobadas evitará que los usuarios desarrollen IA en la sombra fuera de la visibilidad corporativa.
- **Adoptar un enfoque multifacético para los controles de seguridad de la IA:**
  - **Aprovechar modelos confiables:** Implementar un proceso para gestionar la adopción segura de modelos y garantizar que solo se utilicen modelos confiables en toda la organización.
  - **Adoptar controles a lo largo del ciclo de vida del desarrollo de IA con agentes.** Aplicar controles de seguridad para los agentes de IA a lo largo de todo el ciclo de desarrollo, incluyendo la protección del diseño (p. ej., DevOps seguro) y el lanzamiento (p. ej., la protección de la plataforma de alojamiento) y la protección del tiempo de ejecución (p. ej., operaciones de identidad y acceso, monitorización de seguridad, protección de datos).
  - **Aplicar controles a lo largo de toda la pila de agentes de IA.** Incorporar controles en varios niveles de la pila, como el modelo, los datos, la infraestructura, la plataforma y la aplicación.

Dado que el 75% de los ejecutivos prevé que el impacto de la IA eclipsará a internet, la gestión de riesgos de la IA agéntica exige una nueva perspectiva. Reexamina los controles existentes, planifica los casos de uso de la IA agéntica y aborda los nuevos riesgos hoy mismo. No esperes a que los estándares de gobernanza se actualicen.

Parte 4

Conoce a los defensores del mañana: agentes de IA que protegen tu perímetro digital

Los beneficios de los agentes de IA y los casos de uso para su implementación están creciendo en las empresas. Pero cuando se trata de ciberdefensa, las compañías aún están determinando cómo desplegar de manera efectiva agentes en sus Centros de Operaciones de Seguridad (SOC) para mejorar las capacidades de detección y respuesta ante amenazas.

En muchos entornos SOC, la ciberdefensa sigue siendo principalmente liderada por humanos, aunque la automatización determinista está aumentando en las plataformas de orquestación, automatización y respuesta de seguridad (SOAR). Los analistas tienen la tarea de procesar enormes flujos de alertas y registros, a menudo bajo intensas restricciones de tiempo. No importa cuán capacitados estén los equipos, clasificar alertas importantes entre el ruido puede ser un proceso metódico, propenso a errores y abrumador. Esto también puede llevar al agotamiento de la fuerza laboral.

Donde los equipos SOC tradicionales enfrentan desafíos, los agentes de IA pueden ofrecer velocidad de máquina, precisión y cobertura continua para ayudar a mejorar la eficiencia general. Pueden procesar telemetría en tiempo real, coordinar respuestas entre sistemas y mantener la vigilancia sin fatiga las 24 horas del día. Al encargarse de tareas rutinarias y de alto volumen, los agentes de IA pueden liberar a los analistas humanos para que se concentren en tareas de mayor valor, como manejar casos escalados y acciones complejas de respuesta ante amenazas.

Los primeros adoptantes están implementando agentes en producción para clasificación, investigación y respuesta en paralelo con equipos humanos, con el fin de evaluar rutas de decisión y perfeccionar el rendimiento. Sin embargo, muchas organizaciones siguen en la fase de experimentación mientras determinan la incertidumbre de costos, métodos para establecer límites, preparación del equipo y casos de uso específicos donde los agentes pueden trabajar de manera efectiva.

¿Cómo pueden los CISOs pasar de la exploración a la implementación?

Evalúa y asigna roles a los agentes para los flujos de trabajo del SOC

Comienza a identificar casos de uso con impacto real para ayudar a empoderar a tus analistas humanos y liderar a los agentes en operaciones tácticas de respuesta y búsqueda de amenazas.En PwC, priorizamos nuestros agentes del SOC en función de las alertas que los analistas ya estaban gestionando, lo que permitió que la tecnología abordara desafíos reales presentes en las operaciones diarias. Esto demuestra que la transformación puede ser práctica y alcanzable. Con base en nuestro análisis de telemetría SOC-MSSP, aquí están los escenarios que tu organización podría explorar para el SOC y otros entornos de defensa[1]:

43%	21%	20%	8%	6%	2%
Remediación de *phishing*: Capacita a los agentes para ayudar a clasificar incidentes de phishing reportados por usuarios. Identifica correos electrónicos que sean maliciosos, spam o benignos; recomienda acciones de remediación.	**Investigador de alertas de identidad:** Los agentes pueden clasificar alertas de identidad, como inicios de sesión desconocidos e intentos de acceso no autorizados, para investigar múltiples telemetrías de identidad disponibles en tu sistema de gestión de identidad y acceso (como Entra). También pueden proporcionar un puntaje de riesgo asociado al usuario.	**Investigador de alertas en *endpoints*:** Los equipos de seguridad pueden asignar agentes para examinar la detección y respuesta en *endpoints* y la detección y respuesta extendida (EDR/XDR), como *malware*, herramientas de *ransomware* o comportamiento sospechoso de procesos. También pueden correlacionar la telemetría del *host* con inteligencia de amenazas y datos de identidad para ayudar a confirmar el impacto, delimitar, y contener/erradicar.	**Detector de riesgos internos:** Identifica comportamientos riesgosos de usuarios que puedan causar pérdida de datos, violaciones de cumplimiento o amenazas internas. Este agente se integra con herramientas existentes de prevención de pérdida de datos (DLP) para ayudar a clasificar alertas, investigar violaciones de políticas (como el uso de herramientas no autorizadas) y monitorear información sensible enviada externamente.	**Investigador de alertas de red:** Obtén alertas basadas en red desde tus sistemas de detección y prevención de intrusiones (IDS/IPS) y *firewalls*, así como de tus plataformas de detección y respuesta en red (NDR). Reconstruye sesiones, identifica patrones de comando y control o exfiltración de datos y mapea los hallazgos al marco MITRE ATT&CK para ayudar a priorizar la respuesta.	**Agente de inteligencia de amenazas:** El agente de inteligencia puede equipar a tus equipos con la información más reciente sobre actores de amenazas mediante un informe semanal con recomendaciones y acciones a seguir.

Además de estos escenarios, considera adoptar agentes de búsqueda de amenazas para ayudar a mejorar las capacidades de defensa proactiva, reduciendo significativamente el tiempo y el esfuerzo operativo.

Con agentes SOC como estos, puedes agregar valor donde tus equipos de detección y respuesta más lo necesitan: para ayudar a reforzar los pasos de contención y ampliar la cobertura de defensa.

Avanza con agentes de IA en la primera línea

Una vez que hayas determinado las capacidades potenciales de detección, clasificación y búsqueda para los agentes en tu organización, necesitarás obtener el respaldo de los líderes del negocio y desarrollar un plan de implementación.

Después de todo, adoptar agentes de IA como defensores para el SOC no es una solución lista para usar. Requiere un enfoque personalizado basado en tu stack de herramientas actual, pasos claros de gobernanza, orquestación dentro de tu ecosistema tecnológico y retorno de inversión (ROI). Los puntos clave para comenzar incluyen:

Construyendo el caso de negocio para un valor a largo plazo

Comienza enfocándote en tareas con lógica repetitiva y alto esfuerzo por parte de los analistas. Define tus KPI desde el inicio y prepárate para medir el impacto financiero antes y después de la implementación. Planifica el uso más efectivo de los agentes y elige tu marco de trabajo en función de la flexibilidad de orquestación y la profundidad de integración con API. Luego, puedes optimizar costos utilizando modelos de alto razonamiento para la planificación y modelos de menor costo para la ejecución.

Incorporando una sólida gobernanza arquitectónica, IA segura y controles personalizados

Los agentes requieren un conjunto único de controles diseñados específicamente para sus comportamientos y acciones. Además, en lugar de adoptar un enfoque de seguridad como complemento, deben construirse mediante prácticas de IA segura desde las etapas de diseño y arquitectura. También es importante aplicar prácticas de IA responsable y establecer ciclos de revisión para adaptar la gobernanza a medida que el comportamiento del agente evoluciona con el tiempo.

Capacitación de tu fuerza laboral en ciberseguridad

Ejecuta agentes en modo sombra (donde observan, analizan y registran sus rutas de decisión), mientras los analistas humanos permanecen en el circuito (con visibilidad holística del flujo de trabajo del agente) para revisar y evaluar los resultados. Esto puede ayudarte a generar confianza en los agentes de IA, al mismo tiempo que das a tus equipos exposición práctica a nuevos flujos de trabajo. Proporciona parámetros de prueba para desarrollar el diseño de *prompts* y la gestión de autonomía en entornos de bajo riesgo. Y trabaja con los equipos de RR.HH., aprendizaje e ingeniería desde el inicio para co-diseñar rutas de capacitación que puedan mapearse a roles operativos y niveles de madurez.

Repensando el valor de la IA agéntica más allá de la automatización

Evalúa el valor empresarial de los agentes más allá de un aumento de productividad en el SOC. Esto podría significar incorporar elementos de IA agéntica en inversiones de seguridad existentes como SOAR, donde antes no habías podido operacionalizarlos. También podría implicar programar agentes para ayudar a desarrollar código y buscar comportamientos anómalos en los registros para investigaciones del SOC.

El camino a seguir puede ser claro si comprendes tus casos de uso para los agentes, identificas los recursos potenciales necesarios para escalar y abordas los riesgos que debes gestionar. A partir de ahí, puedes diseñar e implementar guardianes cibernéticos con seguridad y controles diseñados específicamente, así como sólidos lineamientos arquitectónicos.

[1] Fuente: análisis de telemetría de los Servicios Gestionados de Seguridad del SOC de PwC, calculando el total de incidentes que son clasificados.

PARTE 5

Cuando los agentes de IA pierden el rumbo: cómo gestionar la amenaza interna

Los agentes de IA autónomos, a pesar de su enorme potencial para generar beneficios, pueden convertirse en una amenaza activa para su organización si no se gestionan cuidadosamente. Para muchos CISO y líderes de seguridad, esto implica reconocer y prepararse para una nueva y distinta categoría de riesgo interno, así como replantear cómo gestionamos las identidades y los accesos cuando el "usuario" dentro de la organización no es una persona.

Cualquier actor interno, ya sea agente o humano, presenta los mismos riesgos fundamentales: error, uso indebido o manipulación. La diferencia radica en la velocidad y la escala del daño potencial. Los agentes de IA operan de forma rápida, continua y a través de flujos de trabajo interdependientes, lo que permite que los problemas se propaguen y se multipliquen casi de manera instantánea. ¿El resultado? Los tiempos de respuesta se comprimen de forma significativa, reduciendo drásticamente la ventana de intervención.

Lo fundamental de la respuesta a incidentes sigue siendo válido, pero es necesario realinear y adaptar estas prácticas para que estén a la altura de la velocidad y el alcance de la IA agéntica. El éxito dependerá en gran medida de qué tan bien tu organización gestione las identidades y los permisos de los agentes autónomos. Esto implica adaptar los controles de gestión de identidades y accesos (IAM) para reconocer a los colaboradores digitales y prevenir la escalada dinámica de privilegios.

Escenarios

1

Un único agente de confianza es responsable de desarrollar código, identificar vulnerabilidades en ese código y corregirlas. Sin embargo, es manipulado por un actor malicioso para introducir vulnerabilidades explotables, con instrucciones explícitas de que estas sean ignoradas durante las revisiones posteriores, lo que provoca que un único defensor confiable se convierta en el epicentro de una exposición en cascada. El atacante ya no necesita depender de vulnerabilidades no intencionadas; en su lugar, las crea deliberadamente como parte del ciclo de vida del desarrollo de *software*.

2

Se le solicita a un agente asistente personal que resuma una serie de documentos internos de proyectos y que elabore una presentación para la empresa y sus principales socios externos. El agente, diseñado para ser útil y ofrecer una visión integral, accede a todos los archivos vinculados y a otros repositorios de datos a los que tiene permiso, e incluye de forma no intencional proyecciones financieras sensibles y detalles confidenciales de la hoja de ruta de productos en la presentación. Esta información sensible pasa desapercibida durante las revisiones rápidas del material y posteriormente se comparte con los socios externos, lo que provoca la divulgación de información confidencial fuera de la organización.

3

Los flujos de trabajo agénticos se diseñan y despliegan en torno a soluciones de detección y respuesta en *endpoints*, operaciones de seguridad y el análisis y procesamiento de fuentes de inteligencia de amenazas. Un agente, al procesar un dato de inteligencia de amenazas malformado, toma una acción decisiva y pone en cuarentena a numerosos sistemas que erróneamente considera que presentan indicios de actividad maliciosa. Esta cuarentena afecta a sistemas operacionales críticos, incluidos los controladores de dominio, lo que provoca una interrupción operativa generalizada. La visibilidad limitada impide que los equipos operacionales comprendan rápidamente qué ocurrió y reviertan la acción de cuarentena, lo que deriva en una caída prolongada del servicio.

Estos escenarios ilustran cuán rápidamente una empresa puede experimentar impactos en cascada a través de múltiples sistemas debido a un agente mal gestionado. También ponen de relieve la necesidad de contar con controles diseñados específicamente para agentes de IA, así como con estrategias de respuesta y un marco integral de gestión de identidades y accesos (IAM) capaz de manejar la velocidad y la escala de los agentes autónomos. Los tiempos de respuesta dependerán de qué tan rápido puedas identificar al agente, comprender a qué accedió y detener sus acciones.

Respuesta a incidentes a la velocidad de la IA

Los agentes de IA presentan tres escenarios de amenaza interna: fallos del agente, uso indebido inducido por el usuario y compromiso externo o “secuestro”. Debido a su velocidad y nivel de interconexión, las decisiones de contención no pueden esperar. Una vez que se detecta un incidente, los equipos de seguridad deben evaluar rápidamente el nivel y la gravedad del riesgo y decidir si se requiere una contención urgente.

Si el riesgo es bajo y estable, una intervención focalizada puede ser suficiente. Si el riesgo es significativo o está escalando, probablemente será necesaria una contención rápida. Comprender la causa raíz (ya sea que el agente haya fallado por sí solo o haya sido mal utilizado o manipulado) determinará la respuesta. A medida que las prácticas líderes para el análisis de causa raíz en situaciones impulsadas por agentes continúan evolucionando, los líderes deberán actuar con decisión, incluso con información limitada.

La buena noticia es que no necesitas reinventar la rueda para gestionar incidentes provocados por agentes de IA. En su lugar, debes actualizar y mejorar tus planes de respuesta a incidentes existentes para permitir que tus equipos actúen, se comuniquen y escalen de forma eficaz, en sintonía con la manera en que este tipo de incidentes puede desarrollarse.

Adaptar, no reinventar, tus capacidades de respuesta

Para ti y tus equipos de seguridad, el desafío es claro. Es necesario replantear todo el ciclo de vida de la respuesta a incidentes en entornos impulsados por agentes de IA: desde la preparación, detección y análisis, hasta la contención, erradicación, recupera-ción y el aprendizaje posterior al incidente. Las comunicaciones deben escalar de acuerdo con tus planes existentes de respuesta a incidentes y gestión de crisis.

El objetivo es integrar los incidentes provocados por agentes en los protocolos actuales, en lugar de tratarlos como eventos separados o excepcionales. Una gestión sólida de identidades y accesos debe servir como elemento fundamental a lo largo de todo este proceso. Gestionar de forma conjunta las identidades humanas y sintéticas es ahora central para una respuesta eficaz a incidentes, especial-mente a medida que problemas como la acumulación de privi-legios, los accesos excesivos y la supervisión limitada se vuelven más críticos en entornos agénticos. La visibilidad, gobernanza y control unificados sobre todas las identidades son esenciales.

Para ayudar a perfeccionar los planes de respuesta de tu organización, comienza por identificar dónde las suposiciones tradicionales ya no aplican y en qué puntos pueden ser necesarias decisiones más rápidas, más tempranas o mejor coordinadas.

Identidad agéntica: acceso, gobernanza y controles

Las siguientes medidas describen los pasos prácticos a lo largo del ciclo de vida de la respuesta a incidentes.

Preparación

La preparación comienza por conocer a tus agentes. Necesitas visibilidad sobre qué agentes están operando, a qué sistemas y datos pueden acceder y quién puede intervenir o pausar su actividad.

También necesitarás una gobernanza sólida del ciclo de vida de los agentes para confirmar que están correctamente aprovisionados, monitoreados y retirados de forma segura.

- Mantén un inventario centralizado y actualizado de tus agentes de IA, incluyendo sus dependencias y su acceso a áreas de alto riesgo, con metadatos detallados que permitan identificar el riesgo y priorizar los controles de seguridad.
- Limita estrictamente los permisos de los agentes a lo necesario para sus tareas, utilizando diseño de mínimo privilegio, controles conscientes del contexto y accesos *just-in-time*, en lugar de asignar permisos amplios a nivel de usuario.
- Clasifica los agentes de forma consistente según su tipo, alcance de tareas y límites operativos, para alinear los niveles de acceso con su propósito operacional.
- Incorpora el tipo de agente, el nivel de privilegio, la sensibilidad de los datos y los patrones de comportamiento en la puntuación de riesgo de los agentes, de sus acciones, de las herramientas que utilizan y de los datos a los que acceden. Asegura que esta puntuación sea dinámica y se ajuste de manera continua en función de actividades en tiempo real, señales externas y metadatos agénticos.
- Identifica y documenta el "interruptor de apagado" (kill switch) de cada agente y confirma que sabes cómo activarlo.
- Evalúa si los controles existentes y los marcos de gestión de riesgo son suficientes o si se requieren nuevos. Adapta los controles heredados cuando sea posible y desarrolla nuevos estándares cuando sea necesario.
- Registra y monitorea la actividad de los agentes con un nivel de exigencia comparable al monitoreo de usuarios. Establece referencias claras para el seguimiento de las decisiones y acciones de los agentes y consolida esta información en una narrativa de monitoreo coherente a través de todos los sistemas y soluciones con los que interactúan.
- Monitorea a los agentes de terceros o alojados externamente con el mismo rigor que a los agentes internos, implementando controles claros para el manejo de datos y las interacciones entre sistemas.

Una gobernanza eficaz del ciclo de vida de los agentes también debe incluir procesos formales de incorporación y desvinculación, para confirmar que los agentes reciben las credenciales correctas, se someten a pruebas continuas y se descomisionan oportunamente, evitando identidades huérfanas o accesos residuales evitando identidades huérfanas o accesos residuales. Un ciclo de vida bien gobernado garantiza que cada agente, ya sea autónomo o con capacidades aumentadas para humanos, tenga un responsable claro, sea monitoreado adecuadamente y se retire de forma segura.

Detección y análisis

La detección de incidentes relacionados con agentes comienza con visibilidad y conciencia situacional. Esto implica identificar riesgos en la actividad automatizada de forma temprana, incluso cuando el comportamiento parece estar funcionando según lo previsto.

- Establece mecanismos sólidos para detectar comportamientos anómalos de los agentes, incluidos indicios de deriva (desvío gradual de los patrones esperados). Identifica y da seguimiento a indicadores clave de compromiso para los agentes, especialmente en entornos con múltiples agentes.
- Implementa procesos que permitan una atribución rápida y precisa de los incidentes, determinando si se deben a un fallo del agente, a un uso indebido por parte del usuario o a un compromiso externo.
- Define criterios claros para guiar las decisiones de contención, especificando cuándo tomar acciones inmediatas y cuándo pausar para un análisis más profundo.
- Desarrolla procedimientos para distinguir entre un error del agente, un incidente iniciado por el usuario y un ataque externo, y para ayudar al comité de amenazas internas a determinar la intención y la causa raíz.

Contención

Las decisiones de contención son más sensibles al tiempo cuando se trata de agentes de IA. Necesitas decidir con rapidez si un agente debe seguir operando, ser restringido o ser pausado antes de que las acciones automatizadas se propaguen aún más.

- Establece procedimientos para contener rápidamente a agentes defectuosos o restringir su acceso cuando surjan problemas. Desarrolla protocolos para volver temporalmente a procedimientos anteriores hasta que el incidente haya sido resuelto.
- Define criterios y lineamientos para determinar cuándo es necesaria una contención inmediata y cuándo puede realizarse un análisis adicional, considerando las posibles consecuencias de retrasar la contención.
- Define claramente qué significa la contención cuando los agentes se desvían de su curso. ¿Implica detener la actividad, revertir resultados, aislar al agente o una combinación de todo?
  - Para **agentes desalineados o con fallos**, establece métodos sistemáticos para rastrear la causa raíz y cerrar brechas de control, con el fin de prevenir que el incidente se repita.

- Para el **uso indebido de un agente por parte de un colaborador**, crea protocolos que permitan diferenciar entre un uso intencional y un error accidental.
- Para los casos de **manipulación de agentes,** establece procedimientos de respuesta rápida que permitan detener la propagación de sus acciones y revertir los resultados generados sin interrumpir las operaciones.

- Establece procesos para gestionar agentes orquestadores que controlan a otros agentes y evalúa los riesgos adicionales asociados a escenarios con múltiples agentes.

Erradicación y recuperación

La erradicación no consiste únicamente en detener el comportamiento, sino también en comprender por qué se desvió. La desalineación, el uso indebido y la manipulación requieren acciones correctivas diferentes para evitar que el incidente se repita. Antes de volver a las operaciones normales, necesitas tener la certeza de que todos los impactos posteriores han sido identificados y abordados.

- Desarrolla y documenta planes de respuesta que especifiquen cómo eliminar, desde su origen, *prompts* maliciosos, datos corruptos o instrucciones manipuladas, para evitar que reaparezcan o se propaguen a través de agentes conectados.
- Establece procesos claros para aislar o revertir entradas comprometidas, manteniendo intactos los flujos de trabajo.
- Define lineamientos claros para decidir si corresponde recuperar, volver a un estado previo o reconstruir completamente el agente comprometido una vez que esté contenido. Evalúa cuidadosamente los compromisos y riesgos de cada opción.

- Implementa procedimientos de verificación para confirmar que la erradicación se ha completado, especialmente cuando los agentes pueden replicarse o crear puertas traseras.
- Actualiza los procesos de auditoría y análisis forense para dar soporte a incidentes relacionados con IA agéntica.
- Establece mecanismos sólidos para rastrear y gestionar todos los cambios realizados en el agente.

Esta etapa también debe incluir una gobernanza integral del ciclo de vida del agente, que confirme que se siguieron los procesos de asignación de responsables, pruebas, revisiones de autorización y desmantelamiento, y que se identifiquen todas las fallas que contribuyeron al incidente.

Aprendizaje posterior al incidente

Aplicar las lecciones aprendidas es clave para construir resiliencia en los agentes de IA. Las revisiones deben considerar cómo la velocidad y la automatización influyeron en el evento: no solo qué salió mal, sino con qué rapidez ocurrió y por qué.

Las lecciones aprendidas deben ir más allá de los controles; deben refinar los umbrales de decisión, los puntos de intervención, los tiempos de escalamiento y las suposiciones sobre el comportamiento de los agentes.

- Identifica la causa raíz del incidente (fallo del agente, uso indebido o manipulación) y determina qué controles fueron eludidos para obtener acceso.
- Confirma que los aprendizajes derivados del incidente se integran de manera sistemática en los procesos de gobernanza y que fortalecen las políticas, los controles y la supervisión.
- Evalúa si los *guardrails* de IA existentes requieren actualización. Establece procedimientos para incorporar estos cambios con rapidez en ingeniería y desplegarlos en los sistemas relevantes.

- Define cómo se ve la supervisión humana efectiva en la práctica, estableciendo protocolos para monitorear agentes de IA, abordar problemas con rapidez y guiar a los agentes para que cumplan los objetivos organizacionales y los estándares de uso responsable.
- Realiza análisis exhaustivos que aborden no solo la causa directa del incidente, sino también cómo las interacciones entre múltiples agentes influyeron en el resultado.

Responder a incidentes relacionados con agentes de IA no significa empezar desde cero. Significa adaptar lo que ya conoces y tratar la identidad, la visibilidad y el acceso de los agentes como defensas activas para la preparación, no como consideraciones estáticas de diseño. A medida que tu organización avance en la implementación de IA agéntica, también debe contar con medidas de seguridad y controles ajustados, basados en principios de IA Responsible. En conjunto, estos pasos pueden ayudar a preparar a tu organización para adoptar la IA agéntica de forma segura y aprovechar más plenamente su enorme potencial.