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.