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.