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.