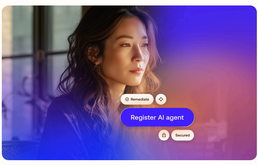
El plan para proteger a una empresa que usa agentes

Descubra las tres preguntas que todo líder de seguridad y TI debe responder antes de que los agentes se descontrolen.



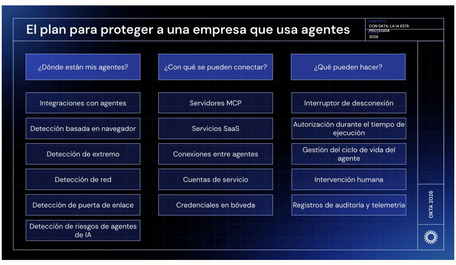
La brecha de identidad en el centro de la seguridad de la IA

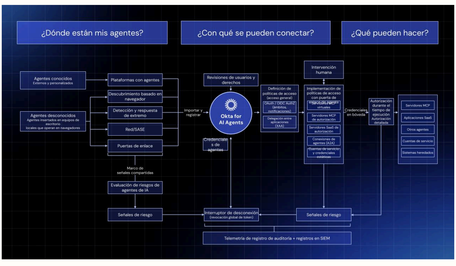
El software solía hacer lo que se le indicaba. Ahora, decide, actúa y se conecta por sí mismo. Los agentes de IA ya están ayudando a los empleados, atendiendo a los clientes y operando en las cadenas de suministro, y están escalando más rápido que la seguridad que los rodea. Durante la última década, las organizaciones fortalecieron la seguridad de la identidad para las personas con sistemas de bóveda, privilegio mínimo y autenticación continua. Sin embargo, el rápido aumento de los agentes de IA está creando una nueva brecha de identidad. Cualquiera puede crear un agente, los agentes pueden generar más agentes y cada uno se conecta a través de aplicaciones, API, herramientas SaaS y sistemas de datos. El resultado son miles de nuevas entidades con acceso privilegiado que operan a la velocidad de las máquinas y suelen escapar a los controles de seguridad actuales.

Por eso, es necesario tratar a los agentes como una identidad de primera clase. Una empresa con agentes segura comienza por establecer niveles claros de responsabilidad y visibilidad antes de que los agentes se descontrolen. Las organizaciones deben ser capaces de responder a tres preguntas:

1. ¿Dónde están mis agentes?
2. ¿Con qué se pueden conectar?
3. ¿Qué pueden hacer?

Estas preguntas definen el plan operativo para proteger a los agentes de IA. Para responderlas, no solo hace falta realizar un inventario; se requieren sistemas, controles de identidad y un modelo de gobernanza adecuados para operar de manera segura una empresa con agentes.





Pregunta 1: ¿Dónde están mis agentes?

La visibilidad es la principal preocupación que nos transmiten nuestros clientes.

Pregunte a su equipo cuántos agentes hay en su entorno. La mayoría no podrá darle una cifra exacta. Por lo general, no se tiene conocimiento ni control de los agentes que los empleados activan en un navegador o que se ejecutan silenciosamente en computadoras de escritorio.

Necesita la capacidad de descubrir agentes sin importar dónde se crearon o se implementaron.

- **Integraciones de plataformas con agentes:** Registre en su proveedor de identidades los agentes de las principales plataformas de terceros y los que se creen en su empresa. Si no puede verlos en el momento de su creación, no podrá controlarlos.
- **Detección basada en navegador:** Descubra agentes en la sombra que operan a través de navegadores y extensiones que se encuentran fuera de su proveedor de identidades. Estos son los agentes que los empleados crean sin pedir permiso.
- **Detección de extremo:** Identifique los agentes que se ejecutan en dispositivos gestionados. Se deberían integrar con cualquier solución que utilice para la gestión de dispositivos móviles o la seguridad de extremos.
- **Detección de red:** Detecte tráfico no autorizado de agente a agente y de agente a recurso en la capa de red. Los agentes se comunican entre sí y con los servicios, y es necesario ver esas conexiones.
- **Detección de puerta de enlace:** Identifique y gestione los agentes de IA no registrados y los clientes de OAuth que interactúan con su API, MCP y puertas de enlace de agentes. Si los agentes llaman a sus API, deben estar autenticados y registrados.
- **Evaluación de riesgos de los agentes de IA:** Supervise e identifique continuamente configuraciones erróneas que hacen que los agentes de IA sean vulnerables a los ataques. Analizar la postura de seguridad de cada identidad de agente permite identificar los riesgos de manera anticipada.

Independientemente de la pila tecnológica que use, necesita recibir señales de todas estas fuentes. Su capa de detección debe funcionar en plataformas, herramientas y equipos. La visibilidad fragmentada es falta de visibilidad.

Pregunta 2: ¿Con qué se pueden conectar?

Una vez que pueda ver un agente, necesita asignar todos los recursos a los que pueda acceder y hacer cumplir las políticas de acceso. Sin un control centralizado sobre las rutas de conexión, un solo agente en riesgo puede abrir el acceso en todo su entorno a la velocidad de las máquinas.

El radio de explosión de un agente comprometido se define por sus conexiones.

- **Servidores y recursos de MCP:** Los servidores de Model Context Protocol brindan a los agentes acceso a herramientas y fuentes de datos, tanto internas (aplicaciones, API, bases de datos, propiedad intelectual) como externas (MCP de terceros como Slack, GitHub y Notion). Su perímetro de seguridad se extiende a todos los recursos a los que su agente puede acceder.
- **Aplicaciones SaaS:** Los agentes se conectan a las mismas herramientas SaaS que los empleados utilizan a diario. La diferencia es que los agentes trabajan mucho más rápido y tienen acceso a más datos. Un agente comprometido puede exponer datos o realizar cambios en todas las aplicaciones SaaS conectadas más rápido de lo que podría hacerlo un humano.
- **Conexiones entre agentes:** Garantice el intercambio inicial y la autorización entre entidades autónomas. Cuando un agente contacta a otro, comienza el movimiento lateral. Ambas partes deben verificar la identidad antes de intercambiar datos o delegar tareas.
- **Cuentas de servicio:** Elimine la proliferación de credenciales estáticas y de larga duración. Estas son las llaves maestras que los agentes reciben de los patrones heredados de máquina a máquina. Cada credencial estática es una puerta a la espera de ser atacada.
- **Credenciales en bóveda:** Proteja y rote los secretos automáticamente. Un token sin rotar es una puerta abierta para los atacantes. Las credenciales deben almacenarse en bóvedas, emitirse dinámicamente y rotarse con frecuencia. Ningún agente debería operar con una credencial que pueda seguir usando después de terminar su tarea.

Todas estas conexiones deben ser registradas en su SIEM. No se puede proteger lo que no se puede ver. Cada conexión de agente, incluida la información sobre a qué accedió, cuándo y con qué credenciales, debe ingresar a su centro de operaciones de seguridad para su supervisión e investigación.

Pregunta 3: ¿Qué pueden hacer?

Saber dónde están los agentes y con qué pueden conectarse no es suficiente si no puede controlar e interrumpir lo que hacen. Cuando un agente comienza a exponer datos o a generar procesos no autorizados, es necesario responder rápidamente.

- **Interruptor de desconexión:** Si un agente se desvía de su misión prevista o accede inesperadamente a datos confidenciales, o si se detecta una amenaza, usted debe poder revocar el acceso de inmediato en todos los sistemas para contener el riesgo.
- **Autorización durante el tiempo de ejecución:** Autorice a los agentes en función de lo que intenten lograr en tiempo real. Evalúe el contexto, la secuencia y el volumen. Una consulta de diez registros de clientes no es lo mismo que una consulta de diez mil. Detecte ataques de inyección de secuencias de comandos y aplique políticas a nivel de las herramientas antes de que se ejecuten las acciones.
- **Gestión del ciclo de vida del agente:** Los permisos de agente que tienen sentido el primer día rara vez tienen sentido el día noventa. Revise siempre el acceso para aplicar el principio de privilegio mínimo, automatizar las certificaciones y revocar el acceso de inmediato cuando se desactiven los agentes o cuando los empleados dejen la empresa.
- **Aprobaciones con intervención humana:** Exija aprobación humana para acciones confidenciales o potencialmente riesgosas de los agentes. Evite operaciones destructivas, el acceso masivo a datos o la escalada de privilegios de los agentes.
- **Registros de auditoría y telemetría:** Cada acción del agente debe ser registrada y enviada a su SIEM. Cada llamada a una herramienta, cada decisión de autorización y cada intento de acceso. La aplicación de medidas en tiempo de ejecución y los interruptores de desconexión solo funcionan si se tiene visibilidad total.

La base está en el plan

Las tres preguntas sobre dónde están mis agentes, con qué se pueden conectar y qué pueden hacer no son hipotéticas. Son el estándar mínimo necesario para utilizar agentes de IA en la etapa de producción.

Las organizaciones que no pueden responderlas están operando a ciegas. Y, cuando la junta directiva pregunte, cuando los auditores pregunten o cuando haya una filtración y el regulador pregunte, la respuesta no puede ser "No lo sé".

Las empresas que están a la vanguardia ya se están ocupando. No esperan a que el primer incidente las obligue a actuar. Ya tratan a los agentes como identidades de primera clase. Están incorporando la detección, el cumplimiento y la gobernanza en sus implementaciones de agentes desde el primer día. Están respondiendo a las tres preguntas antes de que sus agentes se descontrolen.

El plan no es realizar una auditoría una sola vez. Responder a las tres preguntas es una disciplina operativa constante a medida que su población de agentes crece de docenas a miles. Pasó una década desarrollando la seguridad de la identidad para seres humanos. No permita que los agentes tiren todo por la borda.

Okta diseñó este plan basado en el trabajo continuo con empresas líderes que protegen agentes de IA a gran escala. Descubra cómo Okta Platform lo implementa en okta.com/ai-agents.