Riesgo, seguridad y agentes

Cómo se protege la IA, qué puede hacer un agente y con qué proveedores.

> *«El modelo no es una frontera de seguridad. Un agente que actúa necesita identidad propia, permisos mínimos y un interruptor de parada probado; y la responsabilidad no se externaliza al proveedor.»*

QUÉ ES

Qué cambia con la IA

La IA abre ataques nuevos: instrucciones ocultas en el contenido que lee, herramientas manipuladas, fuga de datos. Los agentes, además, actúan. SEVEN-G parte de que el modelo puede ser engañado: el contenido externo no es de fiar, cada acción necesita una intención autorizada y siempre se puede parar.

Implicación: ¿Cuántos agentes de su compañía mueven dinero, datos o sistemas sin que una persona lo valide?

Riesgo: Un agente engañado ejecuta acciones reales con los permisos que se le dieron.

QUÉ PROPONE

Autonomía con límites

La autonomía se fija en cuatro niveles, de solo sugerir a actuar solo. Todo agente tiene identidad, permisos mínimos, límites, registro de acciones y un interruptor de parada probado sin desplegar código. Una persona valida lo que no se puede deshacer. El tráfico a los modelos pasa por una pasarela corporativa.

Implicación: ¿Se ha probado el interruptor de parada de sus agentes, o solo existe en el diseño?

Riesgo: Sin interruptor probado, un agente que falla sigue actuando hasta que alguien lo nota.

CÓMO SE COMPRUEBA

Probar como el atacante

Un control crítico sin diseñar bloquea la construcción; sin probar, la puesta en uso. Quien construye no hace sus propias pruebas adversarias. Con agentes que actúan, cada despliegue pasa por un banco de pruebas corporativo. Con proveedores: salir antes de entrar, y probar con datos internos también es usar.

Implicación: ¿Qué proveedor de IA tiene hoy datos de su compañía sin evaluación previa, aunque sea en pruebas?

Riesgo: Un proveedor sin evaluar puede usar sus datos o dejar a la compañía sin salida.

CUATRO PREGUNTAS PARA EL CONSEJO

1. **¿Qué puede hacer cada agente sin una persona delante?**
   Inventario de agentes con su nivel de autonomía, sus permisos y su dueño. Que un agente actúe solo exige una decisión del comité de IA.
2. **¿Están probados nuestros controles críticos de IA?**
   Todos los controles críticos probados, el interruptor de parada incluido, y ningún ataque de inyección con éxito en acciones sensibles.
3. **¿Pasa todo el tráfico a los modelos por un punto de control?**
   La pasarela corporativa de IA: filtros, modelos aprobados y registro; ante un fallo, se cierra en las acciones sensibles.
4. **¿Podemos salir de nuestro proveedor de IA más crítico?**
   Exigencia al proveedor según su criticidad, cláusulas sobre los datos y, en los más críticos, un plan de salida probado.

QUÉ FALLA SI NO SE GOBIERNA

QUÉ FALLA	NIVEL	QUÉ HACER
Instrucciones ocultas que engañan a un agente	ALTO	Probar cada agente contra la inyección de instrucciones
Agentes con más permisos de los que necesitan	ALTO	Revisar la identidad y los permisos de cada agente
Fraude por suplantación con voz o vídeo sintéticos	ALTO	No aprobar ningún pago solo por voz o imagen
Interruptor de parada que nadie ha probado	ALTO	Probar el interruptor de parada de cada agente
IA incluida en productos de terceros sin evaluar	MEDIO	Desactivar la IA incluida hasta evaluarla
Datos internos en pruebas con proveedores	MEDIO	Evaluar al proveedor antes de cualquier prueba
Dependencia de un solo proveedor de modelos	MEDIO	Preparar y probar el plan de salida del proveedor

PARA SEGUIR LEYENDO

Seguridad de IA y agentes
Documento 35 · amenazas, autonomía y controles

Diseño de seguridad
Plantilla editable

Terceros y proveedores de IA
Documento 36 · exigencia, datos y salida

OWASP Top 10 para aplicaciones agénticas 2026
OWASP GenAI Security Project (Agentic Security Initiative), OWASP Foundation · 09-12-2025

Fernando García Varela · CC BY 4.0 · Versión 0.x, en revisión

Ayuda genérica y gratuita, sin revisión jurídica para ningún caso concreto: cada organización la adapta, identifica la normativa que le aplica y responde de su cumplimiento. No es asesoramiento ni certificación. En la máxima medida permitida por la ley, el autor no asume responsabilidad por los efectos de su aplicación.