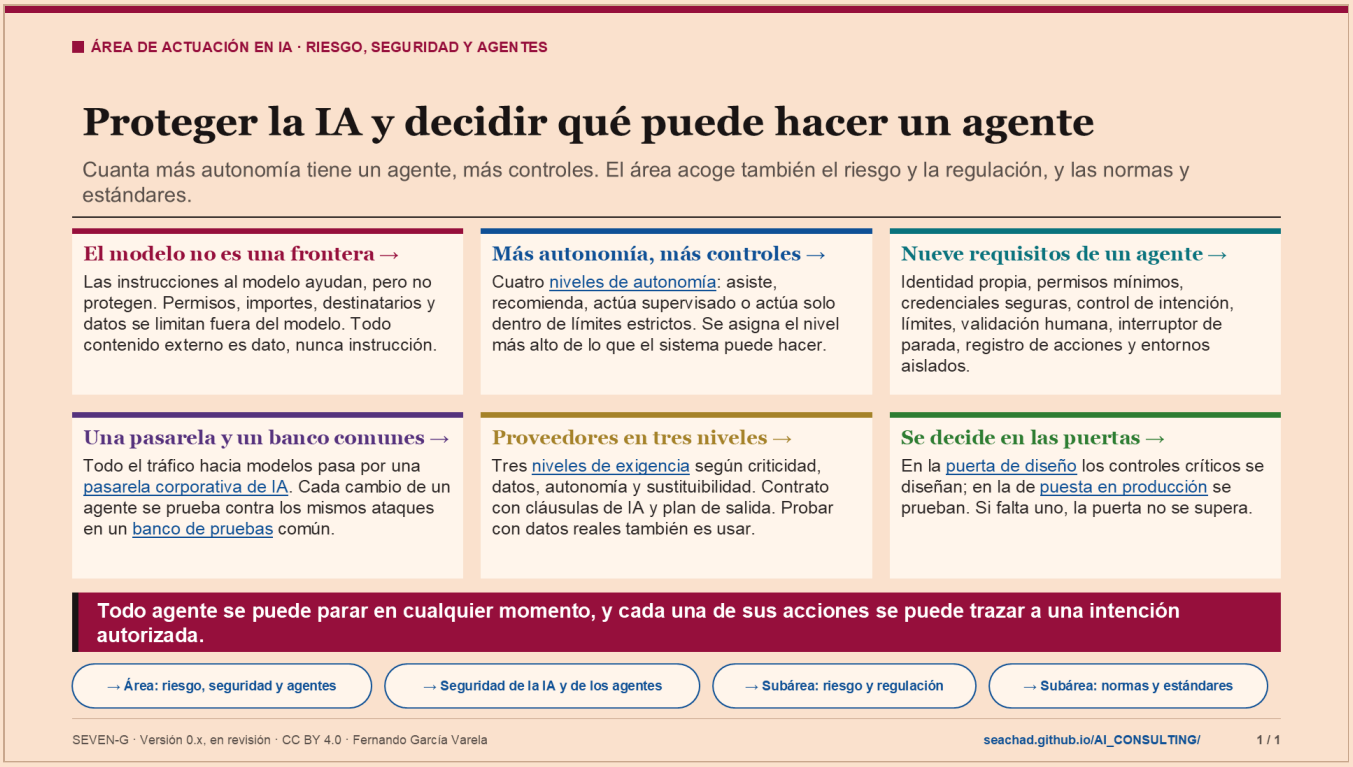

Notas

Esta slide resume el área de actuación Riesgo, seguridad y agentes de SEVEN-G. Responde a tres preguntas: cómo se protege la IA, qué puede hacer un agente y con qué proveedores. Un agente es un sistema de IA que ejecuta acciones sobre otros sistemas, datos o personas mediante herramientas y permisos. Ya no solo responde: paga, escribe, comunica o cambia datos en nombre de la compañía. Por eso su seguridad es un asunto de la dirección, no solo del equipo técnico.

Los seis bloques, uno a uno:

- El modelo no es una frontera de seguridad. Las instrucciones que se dan al modelo ayudan, pero se pueden saltar. Un atacante puede esconder instrucciones en un correo, un documento o una web que el sistema lee: es la inyección de instrucciones. Por eso los límites que importan se aplican fuera del modelo.
- Más autonomía, más controles. SEVEN-G los llama A0 a A3. En A0 el sistema informa o resume y la persona decide. En A1 propone y la persona valida cada acción. En A2 actúa dentro de límites y la persona supervisa. En A3 ejecuta secuencias sin revisión individual, y la persona fija los límites y puede pararlo. Se asigna el nivel más alto de lo que el sistema puede hacer, no el de su uso habitual. Un sistema A3 lo decide el comité de IA en las puertas de diseño (G4) y de puesta en producción (G5).
- Nueve requisitos de un agente. Los cumple todo sistema A2 o A3 y, en lo que corresponda, A1. Destacan tres: identidad propia, distinta de la de cualquier empleado y con un responsable humano; control de intención, que comprueba fuera del modelo que cada acción encaja con lo autorizado; e interruptor de parada, que detiene el agente sin desplegar código y revoca sus credenciales.
- Una pasarela y un banco de pruebas comunes. La pasarela corporativa de IA es el punto de control por el que pasa todo el tráfico hacia modelos, con filtros, prevención de fuga, cuotas y registro. El banco de pruebas corporativo de agentes es una biblioteca versionada de ataques simulados. Sin ellos, la seguridad depende de lo que recuerde cada equipo y nadie puede decir al consejo si los controles funcionan.

- Proveedores en tres niveles. La mayor parte de la IA de una compañía depende de terceros. Cada relación se clasifica en un nivel de exigencia, de N1 a N3, según su factor más exigente. El contrato asegura, entre otras cosas, que los datos de la compañía no entrenan modelos de otros. Y probar una herramienta externa con un documento real ya es usarla: la evaluación del proveedor llega antes que los datos.
- Se decide en las puertas. Un control crítico no diseñado bloquea la puerta de diseño (G4); no probado, bloquea la de puesta en producción (G5). Sobre ellos no se admite continuar con condiciones. Quien construye el agente no hace sus pruebas adversarias antes de la puesta en producción.

Por qué importa. Un error de un asistente es una respuesta mala. Un error de un agente puede ser un pago, un mensaje a un cliente o un dato borrado. Fijar la autonomía, los límites y la parada antes de producción convierte la confianza en la IA en una regla que se aprueba, se prueba y se audita.

Pregunta para el aula: si mañana un agente de su compañía empezara a actuar de forma extraña, ¿quién podría pararlo y cuánto tardaría?

Para ir más lejos, abra la página del área, el documento 35 (seguridad de IA y agentes), el 36 (terceros y proveedores) y la plantilla de diseño de seguridad (P18).

Una precisión sobre el área. Desde que el área se llama «Riesgo, seguridad y agentes», acoge dos subáreas que conservan su página: «Riesgo y regulación» (qué riesgo se asume y qué norma aplica a cada sistema) y «Normas y estándares» (qué cubre SEVEN-G de ISO/IEC 42001, del NIST y del Reglamento Europeo de IA). La seguridad es la parte técnica de un mismo asunto: el riesgo se valora, la norma lo exige y el control lo reduce.

TÉRMINOS DE ESTA SLIDE

- **Niveles de autonomía** (A0–A3): Escala de cuánto actúa un sistema de IA por su cuenta. A0: solo informa; A1: propone y una persona valida cada acción; A2: actúa dentro de límites y una persona supervisa; A3: actúa solo en un ámbito definido. Cuanta más autonomía, más controles: un agente A2 o A3 con efecto sobre terceros es Enterprise. El enlace abre la escala en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#niveles-de-autonomia
- **Pasarela corporativa de IA** (SEG-26): Punto de control común, fuera del modelo, por el que pasa todo el tráfico hacia modelos de IA, propios o de terceros: filtros de entrada, de salida y de contexto, prevención de fuga, cuotas, modelos aprobados y registro; para agentes, el control de cada acción frente a una intención autorizada. Tiene la misma configuración medida para todos los casos de uso. El enlace abre el control en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#cod-seg-26
- **Banco de pruebas de agentes** (SEG-27): Biblioteca común y versionada de ataques simulados, sacada del modelo de amenazas y de los incidentes propios, que prueba agentes y asistentes por todos sus canales, herramientas y conectores. En los agentes que actúan se ejecuta en cada despliegue, y si un escenario crítico tiene éxito no se despliega. El enlace abre el control en el documento 35. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html#cod-seg-27
- **Niveles de exigencia a terceros** (N1–N3): Grado de diligencia, contrato, seguimiento y plan de salida que se exige a cada relación con un proveedor de IA: estándar, reforzado o crítico. Lo fija el factor más alto entre criticidad, datos, autonomía y facilidad de sustitución. Así el esfuerzo se concentra en los proveedores de los que más depende la compañía. El enlace abre los niveles en el documento 36. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html#niveles-de-exigencia-n1-n3
- **Puerta de diseño** (G4): Puerta que aprueba el diseño completo antes de construir: arquitectura, datos, seguridad, supervisión humana, reversión y adopción. Comprueba que los controles están pensados desde el principio. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-5-g4-diseno

- **Puesta en producción** (G5): Puerta que autoriza a usar el sistema de IA con usuarios, clientes o datos reales. Exige la validación y el piloto superados, el plan de reversión probado y la firma de varios niveles (negocio, tecnología, riesgos). Nada entra en producción sin esta firma. El enlace abre sus criterios en el documento 21. https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/21_SEVEN-G_Criterios_de_gate_y_auditoria.html#6-6-g5-puesta-en-produccion

FUENTES Y ENLACES

- **Área: riesgo, seguridad y agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/seguridad.html
- **Documento 35 · Seguridad de IA y agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html
- **Documento 36 · Terceros y proveedores de IA** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/36_SEVEN-G_Terceros_y_proveedores_de_IA.html
- **Glosario (documento 02) · Autonomía de sistemas y agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/02_SEVEN-G_Glosario_y_convenciones.html#4-11-autonomia-de-sistemas-y-agentes
- **OWASP Top 10 para aplicaciones agénticas** https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
- **Seguridad de la IA y de los agentes** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/35_SEVEN-G_Seguridad_de_IA_y_agentes.html
- **Subárea: riesgo y regulación** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/riesgo.html
- **Subárea: normas y estándares** https://seachad.github.io/AI_CONSULTING/SEVEN-G/html/es/areas/normas.html

AVISO LEGAL

Aviso legal: ayuda metodológica genérica y gratuita, «tal cual» y con fines informativos; no es asesoramiento jurídico ni está revisada jurídicamente para ningún caso. Cada organización la adapta y es responsable de su revisión legal y de su cumplimiento regulatorio; puede no recoger cambios normativos recientes. En la máxima medida permitida por la ley, el autor no asume responsabilidad alguna por los efectos de su aplicación. No otorga certificación de ningún tipo. Detalle en el documento 93.

CC BY 4.0 (contenidos) · MIT (código) · Fernando García Varela · seachad.github.io/AI_CONSULTING/