

Lista de control de seguridad de agentes de IA

Catorce controles antes de permitir un agente con herramientas en una operación en vivo. Alineada en espíritu con el OWASP Top 10 para aplicaciones de LLM, sin fingir que una lista es una certificación.

ACTUALIZADO 2026-10-22

Úsela antes de que un piloto toque una cola real, un expediente de reclamo, un sistema central o un cliente. Es una lista de trabajo para un CISO, un responsable de entrega y la persona dueña de la operación. No es una puntuación para publicar, ni una afirmación de que S·CORP o su firma tiene una certificación.

El marco público es el OWASP Top 10 para aplicaciones de LLM: inyección de instrucciones, divulgación de información sensible, cadena de suministro, agencia excesiva y manejo inadecuado de la salida son los puntos que esta lista vuelve concretos. El Marco de Gestión de Riesgos de IA de NIST (gobernar, mapear, medir, gestionar) es una forma razonable de archivar las respuestas. ISO 27001 y PCI DSS importan solo donde esos regímenes ya aplican al sistema.

Cómo usarla

Responda cada línea para un agente, no para "la IA" en general. Un "no" es una razón para estrechar el piloto, no para esconder la línea. Guarde la lista completa con el registro de cambio del agente.

Controles

- ☐ La operación que el agente puede hacer está escrita en una frase, con un dueño.
- ☐ Las acciones que el agente no debe tomar están listadas: mover dinero, cambiar límites, cerrar un caso, enviar datos fuera de la firma.
- ☐ Cada herramienta está nombrada, con un verbo (leer, redactar, crear) y una lista de parámetros permitidos. El texto libre del modelo no es un parámetro.
- ☐ La identidad de la herramienta no es una cuenta de administrador compartida. Una persona puede ver qué agente llamó a qué sistema.
- ☐ El contenido no confiable (correo, PDF, incidencia, página) no puede, por sí solo, cambiar un parámetro. Se probó una instrucción hostil dentro de un documento.

- ☐ El índice de recuperación coincide con un mapa de datos. Los campos fuera de alcance, incluidos los datos de titular de tarjeta, no están embebidos.
- ☐ Los secretos, los nombres internos de servidores y la lógica de reglas de fraude no están en la instrucción de sistema. Asuma que la instrucción puede citarse.
- ☐ La salida del modelo se trata como no confiable antes de mostrarla a un cliente o escribirla en un sistema de registro.
- ☐ Una persona aprueba cualquier acción que cambia un registro. Quien aprueba ve el nombre de la herramienta y los parámetros, no solo un resumen.
- ☐ Una persona nombrada puede sacar al agente del camino en vivo el mismo día, sin un cambio de código, y la detención queda registrada.
- ☐ Cada unidad de trabajo registra el id del caso, las herramientas llamadas, las clases de datos tocadas, la salida mostrada y la decisión humana.
- ☐ El contrato dice dónde se guardan las instrucciones, quién puede leerlas, cuánto viven las trazas, y que el contenido del cliente no se usa para cambiar un modelo salvo acuerdo.
- ☐ La conservación de instrucciones, incrustaciones y trazas no es más larga que la de la operación, y la supresión tiene un dueño.
- ☐ Alguien muestrea transcripciones contra el alcance escrito, en un calendario que el operador reconoce. El método de la muestra está escrito.

Lo que esta lista no hace

- No le dice qué modelo comprar.
- No clasifica su sistema bajo el Reglamento de IA de la UE. Eso es un juicio jurídico.
- No reemplaza una prueba de penetración de las aplicaciones de alrededor.
- No es un caso y no cita resultados de un cliente.

S-CORP — Engineered for outcomes. Built for trust.

s-corp.app · Lista de control de seguridad de agentes de IA