# 5 Reglas de Seguridad para tus Agentes de AI

CATEGORÍA · SEGURIDAD Y BUENAS PRÁCTICAS

Lo que debes revisar antes de darle acceso a tu agente a cualquier herramienta.

QUÉ VAS A ENCONTRAR

FICHA DEL RECURSO

# Qué es, para qué sirve y cómo se implementa

**FUNCIÓN**

Lo que debes revisar antes de darle acceso a tu agente a cualquier herramienta.

**PARA QUÉ SIRVE**

Los prompts de extraños corren con tus permisos. Las skills le dan instrucciones a tu agente. Si esas instrucciones vienen de alguien con malas intenciones, tu agente puede hacer cosas que tú no autorizaste: leer archivos, ejecutar comandos, enviar información fuera de tu sistema. Esto se llama prompt injection. OWASP, la organización de referencia en seguridad de software, lo lista como la vulnerabilidad número 1 en aplicaciones de AI.

**DÓNDE APLICARLO**

Cubre: Por qué las skills públicas son un riesgo real; Como auditar lo que tu agente construye; Las 5 reglas que aplico con mis agentes de Claude; El prompt que uso para construir skills propias; Checklist de auditoría en 5 minutos.

**CÓMO IMPLEMENTARLO**

1. 01 / El Riesgo: Las skills son prompts.
2. 02 / Regla 1: Nunca instales skills de repositorios públicos.
3. 03 / Regla 2: Dale a tu agente solo lo que necesita.
4. 04 / Regla 3: Revisa lo que hace, no solo lo que produce.
5. 05 / Regla 4: Aísla a tu agente del resto de tu sistema.
6. 06 / Regla 5: Dale un kill switch antes de darle permisos.
7. Antes De Activar El Agente: Auditoría en 5 minutos

01 / EL RIESGO

# Las skills son prompts.

Los prompts de extraños corren con tus permisos. Las skills le dan instrucciones a tu agente. Si esas instrucciones vienen de alguien con malas intenciones, tu agente puede hacer cosas que tú no autorizaste: leer archivos, ejecutar comandos, enviar información fuera de tu sistema. Esto se llama prompt injection. OWASP, la organización de referencia en seguridad de software, lo lista como la vulnerabilidad número 1 en aplicaciones de AI.

02 / REGLA 1

## Nunca instales skills de repositorios públicos.

Si ves una skill que quieres que tu agente tenga, dale el link. Dile que entienda qué hace y que construya su propia versión. El resultado: una skill que puedes leer y auditar, sin código de extraños en tu sistema.

**EL PROMPT QUE USO**

Mira esta skill: [link]. Entiende qué hace.
Constrúyeme una versión propia que encaje
en como operamos.

03 / REGLA 2

## Dale a tu agente solo lo que necesita.

Si tu agente maneja correos, no necesita acceso a tus archivos. Si procesa documentos, no necesita internet. Cada herramienta extra que le das es una superficie de ataque adicional. Define el acceso mínimo para cada tarea. Agrega permisos solo cuando son estrictamente necesarios.

**TRES PREGUNTAS ANTES DE ACTIVAR**

¿Qué herramientas necesita para esta tarea?
¿Qué datos tiene que leer?
¿A dónde puede enviar información?

04 / REGLA 3

## Revisa lo que hace, no solo lo que produce.

Activa logs. Lee el historial de acciones de tu agente periódicamente. Si algo parece fuera de lugar, investígalo antes de que se convierta en un problema. Un agente que no monitoreas es una caja negra con permisos.

**SEÑALES DE ALERTA**

Accesos a archivos que no esperabas.
Llamadas a herramientas fuera de la tarea.
Outputs con información que no pediste.

05 / REGLA 4

## Aísla a tu agente del resto de tu sistema.

Si tu agente procesa datos sensibles, dale su propia cuenta, su propio entorno y sus propias credenciales. Si lo comprometen, el daño queda contenido. Nunca uses tu cuenta principal ni tu SSO compartido. La separación es lo que evita que un agente comprometido convierta un incidente en una catástrofe.

**SETUP MÍNIMO**

Cuenta dedicada solo para el agente.
Tokens revocables, no contraseñas.
Acceso solo a la carpeta o base de datos que necesita.

06 / REGLA 5

## Dale un kill switch antes de darle permisos.

Define cómo paras a tu agente antes de activarlo. Si se sale de control, no es momento de aprender cómo. Establece límites duros: tiempo máximo por tarea, presupuesto de tokens, número de acciones permitidas en una corrida. Cuando se pasan, el agente se detiene solo.

**LO QUE DEBE EXISTIR**

Comando o botón para detenerlo al instante.
Timeout máximo por tarea.
Alerta automática si gasta más de X tokens.

ANTES DE ACTIVAR EL AGENTE

## Auditoría en 5 minutos

**LAS 5 REGLAS**

Cero skills instaladas desde repositorios públicos
Acceso mínimo definido para cada tarea
Logs activos y revisados con frecuencia
Agente aislado con cuenta y credenciales propias
Kill switch y límites duros configurados
ANTES DE CUALQUIER CAMBIO
Documentar qué permisos tiene el agente hoy
Probar el kill switch al menos una vez
Revisar logs de la última semana

SIGUIENTE PASO

# Aplica esto hoy. Tu agente mañana es más seguro.

directamente con mis agentes. Sin teoría. Solo lo que funciona.

MÁS GUÍAS GRATIS EN 377CREATIVE.COM/RECURSOS