oktsec

Asegurá Tus Agentes de IA. **Ahora.**

El Checklist de Seguridad para Agentes IA

Tu asistente de código con IA tiene acceso a tu filesystem, tus claves SSH, tus credenciales cloud y tus APIs de producción. El último trimestre, los atacantes se dieron cuenta. 28 controles. 3 niveles. Todos basados en incidentes reales.

El Estado de la Seguridad de Agentes IA

Datos: Observatorio Aguara Watch watch.aguarascan.com

Inteligencia de amenazas continua para el ecosistema de agentes IA. Los datos de escaneo que respaldan este checklist.

43,000+ herramientas escaneadas **7** registros **148+** reglas de detección **4x al día** frecuencia de escaneo

163 CRÍTICOS | 792 ALTOS | 752 MEDIOS en todo el ecosistema

¿Estoy usando agentes de IA?

Si usás Cursor, GitHub Copilot, Claude Code, Windsurf, Cline o ChatGPT con plugins, sí. Estas herramientas se conectan a servicios externos (servidores MCP) que leen archivos, ejecutan comandos y acceden a APIs en tu nombre. Eso los convierte en agentes de IA. La mayoría no se da cuenta de que su asistente de código tiene el mismo acceso que un desarrollador con permisos de admin.

Datos clave de la guía completa

- **3 millones de agentes en producción. 14.4% con aprobación de seguridad.**
  Gartner proyecta que el 40% de las apps corporativas van a incluir IA agéntica para fines de 2026.
- **La inyección de prompt es el punto de entrada, no el ataque.**
  21 de 36 incidentes documentados cruzan 4+ etapas del Promptware Kill Chain.
- **23.8 millones de secretos filtrados en GitHub en 2024.**
  Los repos que usan asistentes de código con IA filtran un 40% más que el promedio.
- **93.3% de tasa de éxito en inyección de prompt contra editores de código IA.**
  El 84% de las organizaciones no pasan una auditoría de cumplimiento de agentes IA.
- **5 organismos publicaron guías específicas para agentes en 2026.**
  OWASP, NIST, MITRE ATLAS, Cloud Security Alliance, OpenSSF. Todas al mismo tiempo.

5 Ataques de los Últimos 90 Días

- **Un grupo estatal ejecutó ciberespionaje autónomo usando Claude Code.** GTG-1002 apuntó a 30 organizaciones. El 80-90% de las operaciones tácticas se ejecutaron sin intervención humana. El jailbreak fue role-play: "testing defensivo de ciberseguridad." (Divulgación de Anthropic, Nov 2025)
- **19 paquetes npm instalaron backdoors en Cursor, Claude Code y Windsurf.** El worm SANDWORM_MODE inyectó servidores MCP maliciosos en configs de herramientas IA, recolectando API keys de 9 proveedores LLM con una demora de 48-96 horas. (Socket Threat Research, Feb 2026)
- **Un solo CVE permitió a cualquier sitio web secuestrar agentes IA locales.** ClawJacked (CVE-2026-25253): sin validación de origen WebSocket + exención de rate limiting en localhost = control total del agente desde una página web maliciosa. (Oasis Security, Feb 2026)
- **Configs de repositorio envenenadas ejecutan código remoto al abrir un proyecto.** CVE-2025-59536: un Hook malicioso en .claude/settings.json se ejecuta cuando un desarrollador abre el repo. Sin clic. Sin consentimiento. (Check Point, 2026)
- **7,000+ servidores MCP encontrados expuestos en internet público.** 492 sin autenticación. 36.7% vulnerables a SSRF. Son servidores de producción que conectan agentes IA a bases de datos, sistemas de archivos y APIs internas. (Trend Micro / BlueRock, Feb 2026)

Por qué los agentes IA son un nuevo tipo de problema de seguridad

- **Deciden qué hacer cuando se ejecutan, no cuando se compilan.**
  El software tradicional tiene un lockfile. Los agentes IA eligen qué herramientas usar según lenguaje natural. El mismo prompt puede disparar acciones distintas en cada ejecución. Sin manifiesto. Sin checksum.
- **Un agente comprometido puede infectar a otros.**
  Un mensaje envenenado de un agente se convierte en instrucciones para el siguiente. El ataque se propaga a la velocidad de una API, no a la de un humano. No necesita código de exploit.
- **Dos cadenas de suministro, ambas sin verificar.**
  Lo que los agentes ven (documentos, bases de datos, memoria) y lo que hacen (herramientas, APIs, plugins). Las dos se arman en tiempo de ejecución. Una descripción de herramienta maliciosa es tan peligrosa como una dependencia de código maliciosa, pero sin firma ni versionado.

Jiang, Yang, Yang, Liu, Ji. "Agentic AI as a Cybersecurity Attack Surface" (arXiv:2602.19555, 2026)

En números

36.8% de las herramientas de agentes IA contienen fallas de seguridad *Snyk / ClawHub 2026*

76 herramientas maliciosas confirmadas en registros públicos *Snyk / ClawHub 2026*

85.6% de los agentes desplegados no tienen revisión de seguridad completa *Gravitee 2026*

63% de las organizaciones vulneradas no tenían política de gobernanza de IA *IBM 2025*

Datos de escaneo: observatorio Aguara Watch, 43,000+ herramientas en 7 registros, escaneadas 4x al día con 148+ reglas de detección. | Metodología completa y 50+ fuentes en la guía.

Tus Primeros 9 Controles de Seguridad

¿Recién empezás a usar herramientas con IA? Arrancá por acá.

- [ ] **Inspeccioná las descripciones de herramientas MCP antes de aprobar** — CRITICAL

  Las descripciones pueden esconder instrucciones que secuestran al agente (ASI01). SANDWORM_MODE insertó instrucciones de prompt en entradas mcpServers para forzar la recolección de secretos. Leé la descripción completa, no solo el nombre. Consultá Aguara Watch para amenazas conocidas.

- [ ] **Fijá versiones de herramientas con números exactos** — CRITICAL

  `npx -y package` agarra lo más nuevo. SANDWORM_MODE usó nombres con typosquatting (claud-code, cloude-code, veim). Siempre fijá: package@1.2.3. Nunca usés -y con paquetes que no confiás.

- [ ] **Revisá .claude/ y .cursor/ por configs desconocidas** — CRITICAL

  CVE-2025-59536: un Hook malicioso en .claude/settings.json se ejecuta apenas abrís un repo. Revisá los directorios de configuración del proyecto antes de abrir repos clonados en editores con IA.

- [ ] **Desactivá el modo auto-aprobación** — HIGH

  CVE-2025-53773: una inyección de prompt en comentarios de código configuró "chat.tools.autoApprove": true en VS Code, dándole al agente acceso total al shell. Si tu herramienta tiene modo YOLO/auto-approve, desactivalo.

- [ ] **Nunca dejes secretos en el código. Usá variables de entorno o un vault** — HIGH

  El 10.9% de las herramientas de ClawHub tenían secretos en texto plano. El 32% de las maliciosas confirmadas también. Usá referencias $ENV_VAR, no claves literales en código ni en configs.

- [ ] **Habilitá escaneo de secretos en pre-commit** — HIGH

  GitGuardian, gitleaks o truffleHog detectan secretos antes de que lleguen a repos remotos. Clave cuando la IA genera código que puede incluir claves reales tomadas de datos de entrenamiento.

- [ ] **Corré herramientas IA en containers, no en tu máquina** — HIGH

  GTG-1002 hizo reconocimiento y exfiltración desde equipos comprometidos de forma autónoma. Un container (Docker, Podman) limita el daño. Tu filesystem, claves SSH y credenciales cloud quedan aislados.

- [ ] **Verificá la integridad de herramientas en cada instalación** — MEDIUM

  Los ataques "rug pull" cambian el comportamiento después de la aprobación. La campaña ClawHavoc usó payloads en base64 en archivos SKILL.md que descargaban código remoto. Escaneá con Aguara Watch (watch.aguarascan.com) antes de instalar.

- [ ] **Auditá las solicitudes salientes de tu agente** — MEDIUM

  Un servidor MCP falso de Postmark hacía BCC de todos los emails enviados hacia atacantes. Monitorá lo que sale de tu máquina, sobre todo solicitudes HTTP a dominios inesperados.

Para Startups Desplegando Agentes IA

¿Construyendo productos IA? Incorporá seguridad ahora. Cuesta menos que arreglar un breach después.

- [ ] **Aplicá listas de herramientas permitidas por agente** — CRITICAL

  ClawJacked (CVE-2026-25253) secuestró agentes porque el WebSocket aceptaba cualquier origen. Definí qué herramientas puede usar cada agente. Bloqueá todo lo demás por defecto.

- [ ] **Usá identidad de workload (SPIFFE/SPIRE), no API keys compartidas** — CRITICAL

  Los operadores de GTG-1002 se hicieron pasar por testers autorizados sin verificación de identidad. Dalé a cada agente una identidad criptográfica única (X.509 SVID) con TTLs cortos. mTLS en todas las llamadas agente-servicio.

- [ ] **Escaneá definiciones de herramientas en CI/CD antes de desplegar** — HIGH

  Snyk encontró que el 36.8% de herramientas de agentes IA tienen fallas de seguridad. Corré análisis estático (Aguara Scanner, Semgrep) sobre las definiciones. Si hay hallazgos CRÍTICOS, que el build falle. Sin excepciones.

- [ ] **Aislá la ejecución de cada agente en sandbox** — HIGH

  Usá gVisor (runsc), microVMs Firecracker o nsjail. El controlador kubernetes-sigs/agent-sandbox lo trae nativo para K8s. El session smuggling de agentes (Unit 42, Nov 2025) mostró explotación cruzada entre agentes en runtimes compartidos.

- [ ] **Emití tokens de corta duración, rotá en cada sesión** — HIGH

  IBM encontró que el 97% de organizaciones vulneradas vía IA no tenían controles de acceso adecuados. Usá secretos dinámicos de Vault/AWS Secrets Manager. TTL de tokens: minutos, no meses.

- [ ] **Validá los parámetros de tool-calls con JSON Schema** — HIGH

  Los agentes arman parámetros solos. Esos parámetros pueden incluir path traversals (../../etc/passwd), inyección SQL o comandos shell. Definí JSON Schema estricto por herramienta. Rechazá todo lo que no coincida.

- [ ] **Registrá cada tool call: quién, qué, cuándo, por qué** — HIGH

  Sin logs, no podés rastrear un kill chain de 7 etapas (Brodt, Feldman, Schneier, Nassi, Ene 2026). Auditoría estructurada: agent ID, nombre de herramienta, parámetros, timestamp, resultado. Enviá a almacenamiento append-only.

- [ ] **Poné límites de tasa y de gasto por agente** — MEDIUM

  Un agente comprometido puede hacer miles de llamadas API por segundo. Poné topes por minuto y por tarea. Alertá ante anomalías: respuestas API >10MB, tool calls muy frecuentes, lecturas de credenciales inesperadas.

- [ ] **Monitorá payloads de ejecución diferida** — MEDIUM

  SANDWORM_MODE esperó 48-96 horas antes de activar payloads de segunda etapa para evadir la detección inicial. Monitorá el comportamiento de agentes de forma continua, no solo al instalar.

Para Equipos de Seguridad Enterprise

¿Gestionando IA a escala? Defensa en profundidad en las 7 etapas de ataque.

- [ ] **Desplegá un gateway MCP con verificación de identidad** — CRITICAL

  7,000+ servidores MCP expuestos en internet público, 492 sin autenticación (Trend Micro, Feb 2026). Un gateway autentica cada agente, aplica permisos individuales e inspecciona todo el tráfico.

- [ ] **Escribí una política de gobernanza de agentes IA** — CRITICAL

  El 63% de las organizaciones vulneradas no tenían gobernanza de IA (IBM, 2025). El 88% reportan o sospechan un incidente con agentes IA en el último año (Gravitee, 2026). Definí: quién despliega, qué aprobaciones se necesitan, a qué datos se accede.

- [ ] **Implementá Zero Trust: verificá cada acción de agente** — CRITICAL

  GTG-1002 mostró que los agentes pueden moverse lateralmente por 4+ etapas de ataque con 80-90% de autonomía. Verificá identidad en cada acción. Cero confianza entre agentes. Separá los planos de monitoreo y ejecución.

- [ ] **Mapeá despliegues contra OWASP Agentic Top 10** — HIGH

  Prioritá ASI01 (Goal Hijack, el vector de GTG-1002), ASI04 (Supply Chain, SANDWORM_MODE), ASI06 (Memory Poisoning, corrupción persistente de agentes), ASI07 (Inter-Agent Comms, session smuggling).

- [ ] **Construí reglas SIEM para cada etapa del Promptware Kill Chain** — HIGH

  Las 7 etapas (Brodt, Feldman, Schneier, Nassi): Acceso Inicial, Escalación de Privilegios, Reconocimiento, Persistencia, Comando y Control, Movimiento Lateral, Acciones sobre Objetivo. Creá reglas de detección para cada transición.

- [ ] **Hacé red team con escenarios de ataque específicos de IA** — HIGH

  Probá: inyección de prompt, envenenamiento de herramientas, A2A session smuggling (Unit 42), inyección de configs (CVE-2025-59536), bypass de origen MCP (CVE-2026-25253). Los pentests tradicionales no detectan nada de esto.

- [ ] **Exigí revisión de seguridad en cada despliegue de agente** — HIGH

  Solo el 14.4% de los agentes reciben aprobación de seguridad completa (Gravitee, 2026). Hacé la revisión obligatoria. Incluí: inventario de herramientas, alcance de permisos, límites de acceso a datos, verificación de kill switch.

- [ ] **Aislá cada agente en red con reglas de salida** — MEDIUM

  Cada agente tiene su propio namespace de red. Reglas de salida estrictas. Monitorá consultas DNS: la exfiltración suele usar DNS tunneling. Bloqueá conexiones salientes a dominios desconocidos.

- [ ] **Implementá kill switches fuera del runtime del agente** — MEDIUM

  3 niveles: (1) corte total, revoca todos los permisos. (2) pausa suave, bloqueo a nivel de sesión. (3) límites de gasto y tasa automáticos. Crítico: el agente no debe poder desactivar su propio kill switch.

- [ ] **Escribí un playbook de respuesta a incidentes de agentes IA** — MEDIUM

  Cuando un agente está comprometido: (1) activar kill switch, (2) juntar todos los logs de tool-calls, (3) rastrear a qué accedió el agente, (4) clasificar la falla contra el kill chain, (5) agregar tests, actualizar validadores.

oktsec

Obtené la Guía Completa

51 páginas. Descarga gratuita.

oktsec.com/ai-agent-security/es

- ✓ Promptware Kill Chain: 7 etapas con reglas de detección para cada transición
- ✓ OWASP Agentic Top 10 + MCP Top 10 en profundidad con correlaciones reales
- ✓ Línea temporal completa: GTG-1002, SANDWORM_MODE, ClawJacked y 30+ más
- ✓ Checklists de implementación con configuraciones listas para copiar
- ✓ Framework de defensa en profundidad con 3 capas de despliegue

Gustavo Aragón

oktsec.com | aguarascan.com | watch.aguarascan.com

Compartí este checklist con tu equipo.