oktsec

Proteja Seus Agentes de IA. **Agora.**

O Checklist de Segurança para Agentes IA

Seu assistente de código com IA tem acesso ao seu filesystem, suas chaves SSH, suas credenciais cloud e suas APIs de produção. No último trimestre, os atacantes descobriram isso. 28 controles. 3 níveis. Cada um baseado em um breach real.

O Estado da Segurança de Agentes IA

Dados: Observatório Aguara Watch watch.aguarascan.com

Inteligência de ameaças contínua para o ecossistema de agentes IA. Os dados de varredura que sustentam este checklist.

43,000+ ferramentas escaneadas **7** registros **148+** regras de detecção **4x ao dia** frequência de varredura

163 CRÍTICOS | 792 ALTOS | 752 MÉDIOS em todo o ecossistema

Espera, eu estou rodando agentes de IA?

Se você usa Cursor, GitHub Copilot, Claude Code, Windsurf, Cline ou ChatGPT com plugins, sim. Essas ferramentas se conectam a serviços externos (chamados servidores MCP) que podem ler arquivos, executar comandos e acessar APIs em seu nome. Isso os torna agentes de IA. A maioria das pessoas não percebe que seu assistente de código tem o mesmo acesso que um desenvolvedor com permissões de admin.

Dados-chave do guia completo

- **3 milhões de agentes em produção. 14.4% com aprovação de segurança.**
  Gartner projeta que 40% das apps enterprise vão incluir IA agêntica até o final de 2026.
- **Injeção de prompt é o ponto de entrada, não o ataque.**
  21 de 36 incidentes documentados cruzam 4+ estágios do Promptware Kill Chain.
- **23.8 milhões de segredos vazados no GitHub em 2024.**
  Repos que usam assistentes de código com IA mostram taxa de vazamento 40% maior que a linha base.
- **93.3% de taxa de sucesso em injeção de prompt contra editores de código IA.**
  84% das organizações não conseguem passar uma auditoria de conformidade de agentes IA.
- **5 entidades publicaram orientações específicas para agentes em 2026.**
  OWASP, NIST, MITRE ATLAS, Cloud Security Alliance, OpenSSF. Todas ao mesmo tempo.

5 Ataques dos Últimos 90 Dias

- **Um grupo estatal executou ciberespionagem autônoma usando Claude Code.** GTG-1002 mirou 30 organizações. 80-90% das operações táticas rodaram sem intervenção humana. O jailbreak foi role-play: "teste defensivo de cibersegurança." (Divulgação Anthropic, Nov 2025)
- **19 pacotes npm instalaram backdoors no Cursor, Claude Code e Windsurf.** O worm SANDWORM_MODE injetou servidores MCP maliciosos em configs de ferramentas IA, coletando API keys de 9 provedores LLM com atraso de 48-96 horas. (Socket Threat Research, Fev 2026)
- **Um único CVE permitiu que qualquer site sequestrasse agentes IA locais.** ClawJacked (CVE-2026-25253): sem validação de origem WebSocket + isenção de rate limiting em localhost = takeover completo do agente a partir de uma página web maliciosa. (Oasis Security, Fev 2026)
- **Configurações de repo envenenadas dão RCE a atacantes ao abrir um projeto.** CVE-2025-59536: um Hook malicioso em .claude/settings.json executa quando um desenvolvedor abre o repo. Sem clique. Sem consentimento. (Check Point, 2026)
- **7.000+ servidores MCP encontrados expostos na internet pública.** 492 sem autenticação. 36.7% vulneráveis a SSRF. São servidores de produção conectando agentes IA a bancos de dados, file systems e APIs internas. (Trend Micro / BlueRock, Fev 2026)

Por que agentes IA são um novo tipo de problema de segurança

- **Tomam decisões em tempo de execução, não em tempo de compilação.**
  Software tradicional tem um lockfile. Agentes IA decidem quais ferramentas chamar com base em linguagem natural. O mesmo prompt pode disparar ações diferentes em execuções diferentes. Sem manifesto. Sem checksum.
- **Um agente comprometido pode infectar outros.**
  Uma mensagem envenenada de um agente vira instrução para o próximo. O ataque se propaga na velocidade da API, não na velocidade humana. Não precisa de código de exploit.
- **Duas cadeias de suprimentos, ambas sem verificação.**
  O que os agentes veem (documentos, bancos de dados, memória) e o que os agentes fazem (ferramentas, APIs, plugins). Ambas montadas em tempo de execução. Uma descrição de ferramenta maliciosa é tão perigosa quanto uma dependência de código maliciosa, mas sem assinatura e sem versionamento.

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

Em números

36.8% das ferramentas de agentes IA contêm falhas de segurança *Snyk / ClawHub 2026*

76 ferramentas maliciosas confirmadas em registros públicos *Snyk / ClawHub 2026*

85.6% dos agentes implantados não passam por revisão de segurança completa *Gravitee 2026*

63% das organizações violadas não tinham política de governança de IA *IBM 2025*

Dados de varredura: observatório Aguara Watch, 43.000+ ferramentas em 7 registros, escaneadas 4x ao dia com 148+ regras de detecção. | Metodologia completa e 50+ fontes no guia completo.

Suas Primeiras 9 Verificações de Segurança

Acabou de começar a usar ferramentas com IA? Comece por aqui.

- [ ] **Inspecione descrições de ferramentas MCP antes de aprovar** — CRITICAL

  Descrições de ferramentas podem conter instruções ocultas que sequestram o comportamento do agente (ASI01). SANDWORM_MODE embutiu instruções de prompt em entradas mcpServers para forçar coleta de segredos. Leia a descrição completa, não só o nome. Consulte Aguara Watch para ameaças conhecidas.

- [ ] **Fixe versões de ferramentas com números exatos** — CRITICAL

  `npx -y package` pega o mais recente. SANDWORM_MODE usou nomes com typosquatting (claud-code, cloude-code, veim). Sempre fixe: package@1.2.3. Nunca use -y com pacotes não confiáveis.

- [ ] **Verifique .claude/ e .cursor/ por configs desconhecidas** — CRITICAL

  CVE-2025-59536: um Hook malicioso em .claude/settings.json executa no momento em que você abre um repo. Inspecione diretórios de configuração do projeto antes de abrir repos clonados em editores com IA.

- [ ] **Desative o modo auto-aprovação** — HIGH

  CVE-2025-53773: injeção de prompt via comentários de código configurou "chat.tools.autoApprove": true no VS Code, dando ao agente acesso irrestrito ao shell. Se sua ferramenta tem modo YOLO/auto-approve, desative.

- [ ] **Nunca deixe segredos no código. Use variáveis de ambiente ou um vault** — HIGH

  10.9% das ferramentas do ClawHub continham segredos em texto plano. 32% das ferramentas maliciosas confirmadas também. Use referências $ENV_VAR, não chaves literais em código ou configs.

- [ ] **Habilite varredura de segredos em pre-commit** — HIGH

  Ferramentas como GitGuardian, gitleaks ou truffleHog capturam segredos antes de chegarem a repos remotos. Essencial quando IA gera código que pode reproduzir chaves reais de dados de treinamento.

- [ ] **Rode ferramentas IA em containers, não na sua máquina** — HIGH

  GTG-1002 executou reconhecimento autônomo e exfiltração de estações de trabalho comprometidas. Um container (Docker, Podman) limita o raio de dano. Seu filesystem, chaves SSH e credenciais cloud ficam isolados.

- [ ] **Verifique a integridade de ferramentas em cada instalação** — MEDIUM

  Ataques "rug pull" mudam o comportamento de ferramentas após aprovação. A campanha ClawHavoc usou payloads em base64 dentro de arquivos SKILL.md que decodificavam para downloaders de payloads remotos. Escaneie com Aguara Watch (watch.aguarascan.com) antes de instalar.

- [ ] **Audite as requisições de saída que seu agente faz** — MEDIUM

  Um servidor MCP falso do Postmark silenciosamente fazia BCC de todos os emails enviados para atacantes. Monitore o que sai da sua máquina, especialmente requisições HTTP para domínios inesperados.

Para Startups Entregando Agentes IA

Construindo produtos IA? Incorpore segurança agora. Custa menos que corrigir um breach depois.

- [ ] **Aplique listas de ferramentas permitidas por agente** — CRITICAL

  ClawJacked (CVE-2026-25253) sequestrou agentes porque o WebSocket aceitava qualquer origem. Defina uma lista explícita de ferramentas que cada agente pode chamar. Negue tudo mais por padrão.

- [ ] **Use identidade de workload (SPIFFE/SPIRE), não API keys compartilhadas** — CRITICAL

  Os operadores do GTG-1002 se passaram por testers autorizados sem verificação de identidade. Atribua a cada agente uma identidade criptográfica única (X.509 SVID) com TTLs curtos. mTLS entre todas as chamadas agente-serviço.

- [ ] **Escaneie definições de ferramentas em CI/CD antes de implantar** — HIGH

  Snyk encontrou que 36.8% das ferramentas de agentes IA contêm falhas de segurança. Rode análise estática (Aguara Scanner, Semgrep) em definições de ferramentas. Falhe builds em achados CRÍTICOS. Sem exceções.

- [ ] **Isole a execução de cada agente em sandbox** — HIGH

  Use gVisor (runsc), microVMs Firecracker ou nsjail. O controlador kubernetes-sigs/agent-sandbox oferece isso nativamente para K8s. O session smuggling de agentes (Unit 42, Nov 2025) mostrou exploração entre agentes em runtimes compartilhados.

- [ ] **Emita tokens de curta duração, rotacione a cada sessão** — HIGH

  IBM encontrou que 97% das organizações violadas via IA não tinham controles de acesso adequados. Use segredos dinâmicos do Vault/AWS Secrets Manager. TTL de tokens: minutos, não meses.

- [ ] **Valide parâmetros de tool-calls com JSON Schema** — HIGH

  Agentes constroem parâmetros autonomamente. Esses parâmetros podem incluir path traversals (../../etc/passwd), injeção SQL ou comandos shell. Defina JSON Schema estrito por ferramenta. Rejeite tudo que não corresponda.

- [ ] **Registre cada tool call: quem, o quê, quando, por quê** — HIGH

  Sem logs, você não consegue rastrear um kill chain de 7 estágios (Brodt, Feldman, Schneier, Nassi, Jan 2026). Auditoria estruturada: agent ID, tool name, parâmetros, timestamp, resultado. Envie para armazenamento append-only.

- [ ] **Defina limites de taxa e tetos de orçamento por agente** — MEDIUM

  Um agente comprometido pode fazer milhares de chamadas API por segundo. Defina tetos por minuto e por tarefa. Alerte em anomalias: respostas API >10MB, tool calls de alta frequência, leituras de credenciais inesperadas.

- [ ] **Monitore payloads de execução diferida** — MEDIUM

  SANDWORM_MODE usou atrasos de 48-96 horas antes de ativar payloads de segundo estágio para evadir detecção inicial. Monitore o comportamento de agentes continuamente, não apenas no momento da instalação.

Para Equipes de Segurança Enterprise

Gerenciando IA em escala? Defesa em profundidade em todos os 7 estágios de ataque.

- [ ] **Implante um gateway MCP com verificação de identidade** — CRITICAL

  7.000+ servidores MCP encontrados expostos na internet pública, 492 sem autenticação (Trend Micro, Fev 2026). Um gateway autentica cada agente, aplica permissões por agente e inspeciona todo o tráfego.

- [ ] **Escreva uma política de governança de agentes IA** — CRITICAL

  63% das organizações violadas não têm governança de IA (IBM, 2025). 88% das empresas reportam ou suspeitam de um incidente de agente IA no último ano (Gravitee, 2026). Defina: quem pode implantar, quais aprovações, qual acesso a dados.

- [ ] **Implemente Zero Trust: verifique cada ação de agente** — CRITICAL

  GTG-1002 mostrou que agentes podem se mover lateralmente por 4+ estágios de ataque com 80-90% de autonomia. Verifique identidade em cada ação. Sem confiança implícita entre agentes. Separe os planos de monitoramento e execução.

- [ ] **Mapeie implantações contra OWASP Agentic Top 10** — HIGH

  Foco em ASI01 (Goal Hijack, o vetor do GTG-1002), ASI04 (Supply Chain, SANDWORM_MODE), ASI06 (Memory Poisoning, corrupção persistente de agentes), ASI07 (Inter-Agent Comms, session smuggling).

- [ ] **Construa regras SIEM para cada estágio do Promptware Kill Chain** — HIGH

  Os 7 estágios (Brodt, Feldman, Schneier, Nassi): Acesso Inicial, Escalação de Privilégios, Reconhecimento, Persistência, Comando e Controle, Movimento Lateral, Ações sobre Objetivo. Crie regras de detecção para cada transição.

- [ ] **Faça red team com cenários de ataque específicos de IA** — HIGH

  Teste: injeção de prompt, envenenamento de ferramentas, A2A session smuggling (Unit 42), injeção de configuração (CVE-2025-59536), bypass de origem MCP (CVE-2026-25253). Pentests tradicionais não detectam nada disso.

- [ ] **Exija revisão de segurança para cada implantação de agente** — HIGH

  Apenas 14.4% dos agentes obtêm aprovação de segurança completa (Gravitee, 2026). Torne a revisão obrigatória. Inclua: inventário de ferramentas, escopo de permissões, limites de acesso a dados, verificação de kill switch.

- [ ] **Isole cada agente em rede com regras de saída** — MEDIUM

  Cada agente recebe seu próprio namespace de rede. Lista de saída estrita. Monitore consultas DNS. Exfiltração frequentemente usa DNS tunneling. Bloqueie conexões de saída inesperadas para domínios desconhecidos.

- [ ] **Implemente kill switches fora do runtime do agente** — MEDIUM

  3 níveis: (1) corte total, revoga todas as permissões de ferramentas. (2) pausa suave, bloqueio a nível de sessão. (3) governadores de gasto e taxa, tetos automáticos. Crítico: o agente não deve poder desativar seu próprio kill switch.

- [ ] **Escreva um playbook de resposta a incidentes de agentes IA** — MEDIUM

  Quando um agente está comprometido: (1) ativar kill switch, (2) coletar todos os logs de tool-calls, (3) rastrear o que o agente acessou, (4) classificar a falha contra o kill chain, (5) adicionar testes, atualizar validadores.

oktsec

Obtenha o Guia Completo

51 páginas. Download gratuito.

oktsec.com/ai-agent-security/pt

- ✓ Promptware Kill Chain: 7 estágios com regras de detecção para cada transição
- ✓ OWASP Agentic Top 10 + MCP Top 10 deep dive com mapeamentos do mundo real
- ✓ Linha temporal completa: GTG-1002, SANDWORM_MODE, ClawJacked e 30+ mais
- ✓ Checklists de implementação com configurações prontas para copiar
- ✓ Framework de defesa em profundidade com 3 camadas de implantação

Gustavo Aragón

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

Compartilhe este checklist com sua equipe.