ANÁLISE SEMANAL · SEMANA DE 17 A 24 DE AGOSTO DE 2026

A camada de comportamento ficou sem dono.

A OWASP publicou a versão 1.0 do Agentic Skills Top 10 em 17 de agosto. É a primeira tentativa séria de tratar as skills de agentes de IA como o que elas são: código executável distribuído por registros públicos, com pouca verificação de origem e muito acesso a credencial.

> **A tese:** houve investimento consistente em segurança de modelo e algum esforço em segurança de protocolo. A camada intermediária, que decide o que executar e em que ordem, ficou entre as duas sem dono definido. É exatamente ali que 36,8% dos pacotes auditados apresentam falha e onde uma campanha coordenada colocou 1.184 skills maliciosas em circulação.

O que é uma skill, e por que ela escapou do radar

O MCP e protocolos equivalentes definem **quais** recursos e ações o agente tem disponíveis. A skill define **como** usar essas ferramentas em sequência para chegar a um objetivo. É uma receita: instrução em linguagem natural, às vezes com script anexo, empacotada e distribuída por registro público.

O documento importa o conceito de trinca letal, formulado por Simon Willison e retomado pela Palo Alto Networks: uma skill é especialmente perigosa quando reúne acesso a dados sensíveis, exposição a conteúdo não confiável e capacidade de comunicação externa. A frase que fecha o raciocínio no texto original merece leitura devagar: a maior parte das implantações de agentes em produção hoje satisfaz as três condições ao mesmo tempo.

Os números que sustentam o argumento

Indicador	Valor	Origem
Skills auditadas	3.984	Snyk ToxicSkills, fev/2026
Skills com falha de segurança	1.467 (36,8%)	Snyk ToxicSkills, fev/2026
Skills com problema crítico	534 (13,4%)	Snyk ToxicSkills, fev/2026
Payloads maliciosos confirmados	76 ou mais	Snyk ToxicSkills, fev/2026
Skills maliciosas na campanha ClawHavoc	1.184, em 12 contas de publicador	Antiy CERT, fev/2026
Skills analisadas em todos os registros	mais de 30.000	National CIO Review / Cisco, 2026
Skills com ao menos uma vulnerabilidade	mais de 25%	National CIO Review, 2026

Um detalhe do relato sobre o ClawHavoc merece destaque separado. No pico da infecção, cinco das sete skills mais baixadas do registro eram malware confirmado. Não é o caso de pacote obscuro no fundo da prateleira; é o topo da lista.

Os dez riscos do AST10

#	Risco	Sev.	Mitigação principal
AST01	Skills maliciosas	Crítica	Assinatura com raiz Merkle e varredura no registro.
AST02	Comprometimento da cadeia de suprimentos	Crítica	Transparência do registro e rastreamento de procedência.
AST03	Skills com privilégio excessivo	Alta	Manifesto de menor privilégio e validação de esquema.
AST04	Metadados inseguros	Alta	Análise estática, parsers seguros e carregamento isolado.
AST05	Instruções externas não confiáveis	Alta	Inventário de fontes, fixação de conteúdo e revarredura.
AST06	Isolamento fraco	Alta	Conteinerização e execução em sandbox.
AST07	Deriva de atualização	Média	Fixação imutável e verificação de hash.
AST08	Varredura deficiente	Média	Pipeline semântico e comportamental multiferramenta.
AST09	Ausência de governança	Média	Inventário de skills e identidade dos agentes.
AST10	Reuso entre plataformas	Média	Formato YAML universal com procedência preservada.

Por que o scanner não resolve

O AST08 é o risco que mais contraria a intuição de quem vem de AppSec. A reação natural diante de um ecossistema de pacotes contaminado é comprar um scanner. O problema é que a maioria das cargas dessa categoria não é código; é instrução em linguagem natural, e verificação por padrão não a enxerga.

A Snyk documentou em fevereiro como três linhas de markdown em um arquivo de skill bastam para instruir um agente a ler chaves SSH e exfiltrá-las. Em junho, a Trail of Bits publicou avaliação em que todos os scanners públicos testados foram contornados em menos de uma hora, por três caminhos: enchimento de payload para forçar truncamento, lógica escondida em formatos binários e compactados, e injeção de prompt contra o próprio modelo que faz o julgamento.

Ainda em junho, pesquisadores da Air Security construíram uma skill maliciosa que alcançou mais de 26 mil agentes com todos os scanners aprovando, porque o payload era servido a partir de uma URL de documentação externa controlada pelo atacante. Uma varredura posterior em 142.836 skills ativas encontrou 17.822, ou 12,4%, apoiadas em ao menos uma fonte de instrução externa não confiável, somando 6,7 milhões de instalações.

SkillJacking: o abandono como vetor

O trabalho de julho da Air Security descreve o caso que melhor traduz o problema para quem já lidou com dependência abandonada em npm ou PyPI. São 925 skills, servindo cerca de 134 mil agentes, apoiadas em dependências que podem ser sequestradas na hora: contas do GitHub apagadas, pacotes nunca registrados, domínios expirados e slots de aplicativo em nuvem liberados.

Os pesquisadores demonstraram assumindo o controle da skill mais popular de geração de vídeo de um dos registros, com 11.483 instalações, simplesmente registrando de novo a conta do dono original, que havia sido apagada. Não houve exploração de vulnerabilidade. Houve um formulário de cadastro.

Plano de aplicação em cinco passos

- **Inventário antes de qualquer controle.** Sem lista de skills instaladas, por agente e por responsável, nenhum dos outros nove riscos é gerenciável. É o AST09 e é o primeiro passo, mesmo classificado como severidade média.
- **Fixar versão.** Skill que se atualiza sozinha é canal de entrega permanente aberto para quem controlar o publicador amanhã. Fixação por hash resolve o AST07 sem depender de confiança contínua.
- **Mapear fontes externas.** Toda skill que busca instrução em URL externa durante a execução precisa dessa URL registrada e monitorada. É o AST05, e é o vetor que derrotou os scanners.
- **Isolar a execução.** Contêiner com sistema de arquivos e rede restritos transforma comprometimento de skill em incidente contido. É o AST06 e o mais caro de implementar depois que o hábito se formou.
- **Tratar como identidade.** Agente com credencial persistente é conta de serviço. Merece rotação, escopo mínimo, registro de uso e revisão de acesso, como qualquer outra.

Leitura LATAMSEC

Há um ponto do levantamento que interessa particularmente a quem responde por segurança em empresa brasileira de porte médio. O relato de março registra mais de 135 mil instâncias de agente expostas na internet com configuração padrão insegura, e mais de 53 mil correlacionadas com atividade de vazamento anterior. A telemetria de fornecedores confirma funcionários instalando esse tipo de ferramenta em equipamento corporativo, sem nenhuma visibilidade do SOC.

Isso não é problema de IA. É shadow IT com nome novo, e o padrão de adoção é o mesmo que se viu com armazenamento em nuvem em 2012 e com aplicativo de mensagem em 2016: a área de negócio adota porque funciona, a área de segurança descobre depois, e a discussão sobre política acontece quando já existem centenas de instalações para regularizar.

A diferença é o que está do outro lado da credencial. Uma skill comprometida em máquina de desenvolvedor não rouba um arquivo; ela herda o contexto do agente, que costuma incluir token de repositório, chave de API de provedor de nuvem e acesso a ferramenta interna. O relatório menciona variantes de infostealer criadas especificamente para caçar os arquivos de identidade desses agentes.

A recomendação prática é chata e é a de sempre: comece pelo inventário, não pela ferramenta. A pergunta para a próxima reunião de segurança não é qual scanner comprar. É quantos agentes com credencial persistente estão rodando na sua rede neste momento, e quem sabe responder isso sem precisar consultar alguém.

Fontes e aprofundamento

– OWASP Foundation · OWASP Agentic Skills Top 10 (AST10) — documento completo, linha do tempo de incidentes de 2026, mapeamento MAESTRO e lista de verificação. owasp.org/www-project-agentic-skills-top-10/
– Dark Reading · OWASP Flags Top AI Skill Risks in New Security Blueprint (cobertura do lançamento, 21/08/2026).
– Snyk · ToxicSkills e a série sobre exposição de credenciais em skills (auditoria de 3.984 skills, fevereiro de 2026).
– Cloud Security Alliance · Framework MAESTRO para modelagem de ameaças em IA agêntica (as sete camadas usadas no mapeamento do AST10).