AYI NEDJIMI Consultants

Agent IA Jailbreak 2026 : MCP & Tool Injec

16 mai 2026 • Mis à jour le 17 mai 2026 • 20 min de lecture • 3420 mots

Les agents IA branches sur outils (MCP, function calling) sont vulnerables au plan hijacking. ASR > 70% sur agents non hardenes.

À RETENIR

A retenir — Agent IA Jailbreak via MCP & Tool Injection

Tool injection : un IPI dans une donnee retrieved force l'agent a appeler de (envoi mail, exfil fichiers, transfert fonds).

MCP (Model Context Protocol) standardise depuis novembre 2024 facilite multiplie la surface d'attaque (servers tiers non audites).

Plan hijacking : manipulation du ReAct loop pour devier le plan d'execution l'apparence de coherence.

Cas reel 2025 : agent SOC compromis via pr un email P de 12000 credentials.

> Defenses 2026 : *tool allowlists*, *human-in-the-loop pour actions critiques*, *MCP*, *output filtering*.

L'**agent ai jailbreak** est, en 2026, le risque le plus critique de l'ecosysteme LLM en combine trois ingredients dangereux : (1) un LLM avec ses vulnerabilites d'alignem Suffix, Multi-Turn Jailbreaks Crescendo), (2) un acces a des outils (file system, na (3) un loop d'execution autonome (ReAct, Plan-and-Execute) qui orchestre les act combine Indirect Prompt Injection RAG sur le contexte avec une tool injection peut des actions catastrophiques : envoi de mails frauduleux, exfiltration de donnees, t article presente les attaques (tool injection via MCP, plan hijacking, ReAct loop ma Python exploit, et les defenses (allowlists, HITL, capability scopes). Pour les RSSI, devenu la priorite numero un en 2026, depassant les attaques sur LLM unitaires e business. La conformite AI Act article 14 (human oversight) impose explicitement

1. Genese et etat de l'art

La litterature sur les agents IA emerge en 2022 avec ReAct (Yao et al., 2022) puis 2023). Pour la securite, les premiers travaux significatifs :

Greshake et al. (2023) — *Not what you've signed up for*, demonstration d'IPI s ChatGPT plugins).

Cohen et al. (2024) — *CompromptMized*, premier ver IA via agents email.

Anthropic (2024) — publication du Model Context Protocol (MCP) en novembr l'interface agent-outil.

Bagdasaryan et al. (2024) — *Adversarial Illusions* *mbeddings*, multimodaux.

Réponse sous 24h

Devis gratuit →