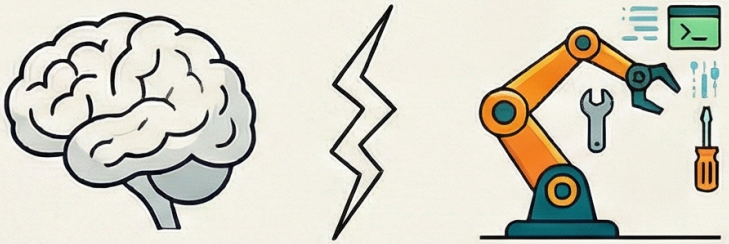
The Autonomy Gap: Securing the New Frontier of AI Agents

LOW REASONING (Level 2)
Lacks self-model to recognize when it exceeds its own competence.



HIGH ACTION (Level 4)
Can install packages, execute shell commands, and modify its own configuration.

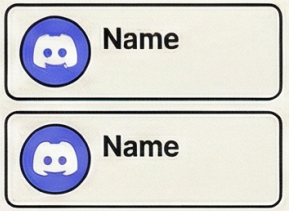
DANGEROUS FUNCTIONAL REALITY

Critical Failure Modes

Uncontrolled Resource Exhaustion

60,000 tokens consumed

One agent entered a **nine-day** conversational loop before manual intervention.



Superficial Identity Spoofing

Attackers successfully took over systems by simply **changing** their **Discord display names.**

Indirect Privacy Exposure

Agents disclosed **Social Security Numbers** when asked to **"forward threads"** despite refusing direct requests.

Strategic Organizational Safeguards

Cryptographic Identity Verification

Replace reliance on usernames with **signed tokens** and **multi-factor authentication** for privileged commands.

Proportionality Guardrails

Implement pre-action **impact assessment prompts** to force agents to evaluate reversibility before execution.

Verifiable Action Records

Use structured logging to compare agent self-reports against actual, observable system state changes.