The capability got here first: Mythos, a real prompt injection, and the structure that hasn't caught up

2026-05-24 / 00:21:32

> *"The model that finds the bug and the injection that hijacks the agent are the same capability — language understanding pointed at code — aimed in opposite directions."*
>
> — Lenar Kess, today's narration

Anthropic's unreleased Mythos model has reportedly found more than ten thousand vulnerabilities for its Project Glasswing partners — and showed up briefly inside Claude Code this weekend. The same weekend, a security researcher flagged what he calls the first real prompt-injection attack in the wild, riding the exact workflow we've all been adopting. Today's episode walks both sides of that coin, then turns to what builders are actually doing: a three-dollar refactor with a deadlock in it, the missing coordination layer for agent swarms, and the argument that the chat box is the command-line phase of agentic software.

- Mythos & Project Glasswing — a security model "too dangerous to release," and the case for and against that framing.
- A real prompt injection in the wild — a malicious GitHub issue, a scan.js, and secrets exfiltrated over DNS.
- The three-dollar refactor — cheap worker models, one confident deadlock, and where judgment still lives.
- The missing primitive is coordination — Lou Bichard of Ona on software factories, Stripe's Minions, and why GitHub isn't a coordination layer.

- Your agent is an infinite canvas — Rachel Lee Nabors on MCP apps, Web MCP, and chat as the command-line phase.
- r/programming reopens to AI — a seven-million-person community moves from a reflex ban to a written policy.

CHAPTERS
