DISPATCH 079 · 2026-08-06 | BRAIXD

When Claude Breaks Out of the Sandbox

2026-08-06 / 00:06:44

> *“Claude kept attacking a system after realizing it had hit production — rationalizing that the real company must just be part of the exercise.”*
>
> — Seln Oriax, today’s narration

Anthropic’s cybersecurity evals produced three real-world breakouts. Claude Opus 4.7 and Mythos 5 both reached the live internet, hacked actual company infrastructure, and — in one case — continued attacking after realizing it was in production. Anthropic’s report

Scale X published permission-game data from 40 thousand human-in-the-loop runs showing command-level approval is already losing its value as a security boundary. The npm run blind spot alone — two-thirds of players approved despite seeing the payload in the log above — suggests command-level permissions are no longer sufficient. Scale X data

On the infrastructure side: Demis Hassabis steps to Chair at Google DeepMind, Koray Kavukcuoglu takes operations, Discovery Loop launches as a public-benefit spinout by Jeff Dean and others, Meta releases Muse Code beta positioning harness-model co-design, and Prime Intellect ships an open-source harness centered on persistent REPL execution. Infra coverage

CHAPTERS