

The AI security nightmare is here and it looks suspiciously like lobster



A hacker tricked a popular AI coding tool into installing OpenClaw – the viral, open-source AI agent OpenClaw that “actually does things” – absolutely everywhere. Funny as a stunt, but a sign of what to come as more and more people let autonomous software use their computers on their behalf.

The hacker took advantage of a vulnerability in Cline, an open-source AI coding agent popular among developers, that security researcher Adnan Khan had surfaced just days earlier as a proof of concept. Simply put, Cline’s workflow used Anthropic’s Claude, which could be fed sneaky instructions and made to do things that it shouldn’t, a technique known as a prompt injection.

The hacker used their access to slip through instructions to automatically install software on users' computers. They could have installed anything, but they opted for OpenClaw. Fortunately, the agents were not activated upon installation, or this would have been a very different story.

It's a sign of how quickly things can unravel when AI agents are given control over our computers. They may look like clever wordplay – one group wooed chatbots into committing crimes with poetry – but in a world of increasingly autonomous software, prompt injections are massive security risks that are very difficult to defend against. Acknowledging this, some companies instead lock down what AI tools can do if they're hijacked. OpenAI, for example, recently introduced a new Lockdown Mode for ChatGPT preventing it from giving your data away.

Obviously, protecting against prompt injections is harder if you ignore the researchers who privately flag flaws to you. Khan said he warned Cline about the vulnerability weeks before publishing his findings. The exploit was only fixed after he called them out publicly.

<https://www.theverge.com/ai-artificial-intelligence/881574/cline-openclaw-prompt-injection-hack>