Putting the Genie Back in the Bottle: Agentic Reverse Engineering of Claude's Security Architecture

.ical (/recon-2026/talk/DZUQYU.ics) (/recon-2026/talk/DZUQYU/feedback/)

2026-06-21 14:30–15:30 18:30-19:30 (UTC), Grand Salon Opera

Language: English

(/media/recon-2026/submissions/DZUQYU/Todd-Manning-REcon-Title-Image_qG_6i7F4tT.webp)

The proliferation of AI agents is quickly becoming one of the foremost concerns of security teams. Engineering teams are clamoring for the increase in velocity afforded by AI coding agents. Non-technical teams have noticed, and employees of all job types are asking for agentic AI tools to facilitate their work. Security teams need to have a clear understanding of how these tools operate, what their security features are, and where the security failures lie. Armed with this knowledge, security teams can enable these new agentic work paradigms while protecting all the things.

This talk presents the complete reverse engineering of Anthropic's Claude Code, Claude Desktop, and Claude Cowork. The recent release of Claude Cowork provides the LLM agent with extraordinary host privileges -- spawning VMs, mounting host directories, taking screenshots, typing into terminals, automating browsers and applications -- all decided by a language model one prompt injection away from hostile intent. We take a look at the two personalities of Claude Cowork. One component of Cowork is Claude Code, running inside a Linux VM using multiple isolation strategies to constrain LLM agent access to user resources, and another is Claude with agentic access to dive the desktop user interface, with capabilities for reading and interacting with anything on the screen.

In this talk, we also present the power of agent-assisted research and development for not only understanding the features and attack surface of these Claude agents, but we demonstrate newly-discovered vulnerabilities in components of Claude. We also identify attack surfaces that in some cases are obvious to see, and other attack surfaces that are completely surprising to discover.

We investigate binaries spanning multiple languages, including Swift, Rust, Go; two JavaScript runtimes; recovering the complete VM hardware configuration from decompiled Swift; the full vsock RPC protocol from a stripped Go guest agent; examine Claude's cloud based and local configuration systems; perform analyses of the Linux VM container isolation strategies; and uncover a hidden

BLE hardware companion protocol that provides auto-approve capabilities (effectively 'dangerous permission mode') for every tool request the model makes. We present confirmed vulnerabilities in multiple subsystems.

Finally, we draw some conclusions about the security architecture of Claude Desktop as a whole, identifying some glaring gaps in which threats the architecture prioritizes, and which seem to have been woefully ignored. We investigate strategies for improving isolation of the agent, and consider where these might fall short.

(/recon-2026/speaker/P83TYB/)

Todd Manning (/recon-2026/speaker/P83TYB/)

I have extensive experience across diverse industry verticals such as automotive, banking, medical, mobile, embedded, industrial control, public utilities, oil & gas, wired and wireless networking, telecommunications, cloud computing, and AI.

As a key member of advanced security research teams at BreakingPoint Systems, Accuvant Labs, Optiv Security, Duo Security, Trend Micro, Atredis Partners, and Together AI, I have successfully delivered hardware and software products, security research, and consulting services to customers and the wider security community.

I attended the conference a couple of times in the 2008-2012 time frame. You don't need me to tell you how great REcon is. It's really great.