ISSUE 1 · AI VULN DISCOVERY IN MEDTECH

Myth about Mythos

First Issue

KEY TAKEAWAYS

- The legacy triage model (a CVE drops, spend two days deciding if it applies, log a row in the cyber risk spreadsheet, repeat ~15-20 times a month) is finished now, not in five years.
- Inbound vulnerability finding volume has gone vertical in the last twelve months because AI agents do a researcher's week of discovery in an hour against any reachable target, reading firmware, SBOM, protocol stack, and public docs simultaneously.
- 'AI found this vulnerability' is not a magic prompt to a magic model; frontier reasoning is the engine but reasoning is not output, so the model is necessary but not sufficient.
- The real machine is a swarm: a recon agent enumerates the target, a fuzzer mutates inputs and logs crashes, a coverage agent unblocks the fuzzer when it stalls on a CRC/ length field/magic byte, a verification agent decides which crashes are reachable, and an orchestrator holds the plan, budget, and success criteria.
- Agents do not share context; they communicate through small structured handoffs and a shared filesystem that is the team's long memory, compacting and re-reading progress files when a context window fills.
- The plumbing is now democratized: Anthropic's Claude Agent SDK (same runtime as Claude Code, with subagent orchestration, session persistence, auto context compaction, 14 built-in tools, MCP) is pip install, so a public-target run costs tens of dollars and a niche embedded device a few hundred, with no nation-state premium left.

> **"It's because AI agents are doing the discovery work that used to take a researcher a week, in an hour, against any target that's reachable."**

Myth about Mythos

The misconception is that the model is the whole show. It is not: frontier reasoning is remarkable but reasoning is not output, and turning it into real discovery needs the swarm around it, the agentic loop, tool integrations, shared filesystem, coverage feedback, and domain skills. Plug a frontier model into a well-built swarm and results are extraordinary; hand it a bare prompt and you get a confident paragraph that misses the bug.

How AI agents are doing vulnerability discovery

Deployment is a swarm, not one model asked to find a bug. A recon agent produces a structured model of the target, a fuzzer generates traffic and logs crashes, a coverage agent diagnoses why the fuzzer is stuck and hands it a structured fix, a verification agent decides reachability, and an orchestrator keeps the plan. They coordinate through handoffs and a shared filesystem that serves as long memory.

Who can do this now

Anyone with a credit card, but without vulnerability-research expertise the results are 'unexploitable' vulnerabilities that still need a pipeline of agentic workflows to weaponize. The Claude Agent SDK and equivalents from OpenAI and Google turned the hard plumbing into a library, collapsing cost and erasing the nation-state premium.

What it actually means

Four consequences in order of worry: volume keeps climbing and is about to inflect again; most of those findings do not matter to a specific product but still must be risk-assessed; the current playbook of manual triage, quarterly pentests, and annual threat models cannot scale and you cannot hire your way out; and the regulatory bar (524B, CRA 14, NIS2 23, IMDRF) is moving slowly but auditors will soon ask how you proved a CVE was not exploitable.

What to actually do

The primitives that make attack cheap make defense cheap; the swarm pattern is symmetric. Teams that point it inward gain a structural advantage by building continuous AI triage, verifying exploitability with artifacts, keeping threat models as persistent assets, and producing regulatory evidence as a byproduct.

What this newsletter will be

Weekly, short, focused on what is actually changing in product security as AI rewrites both offense and defense. Field notes from real assessments, no vendor pitches and no 'rapidly evolving landscape' rhetoric.

Manual triage cadence cited at roughly 15 to 20 CVEs per month per team, maybe more.

Teams reporting 2 to 3x the inbound finding volume they saw a year ago, before the next wave of autonomous tooling.

Discovery that used to take a researcher a week now takes about an hour.

Claude Agent SDK: 14 built-in tools plus MCP support for custom tools.

Run cost: tens of dollars against a public-facing target; a few hundred dollars against a niche embedded device (example given: a hospital MRI scanner exposed through a misconfigured VPN).

The barrier collapsed 'sometime in the last nine months.'

About ELTON. ELTON is the leading AI vulnerability discovery and cybersecurity verification platform for medical devices. ELTON builds a digital twin of each device, continuously identifies new vulnerabilities across the full stack, and uses AI to determine which are actually exploitable, with every finding and every dismissal evidenced for regulatory review. ELTON meets FDA premarket (524B) and postmarket cybersecurity vulnerability testing and management requirements. Read the newsletter and talk to us at **eltoncyber.com**.