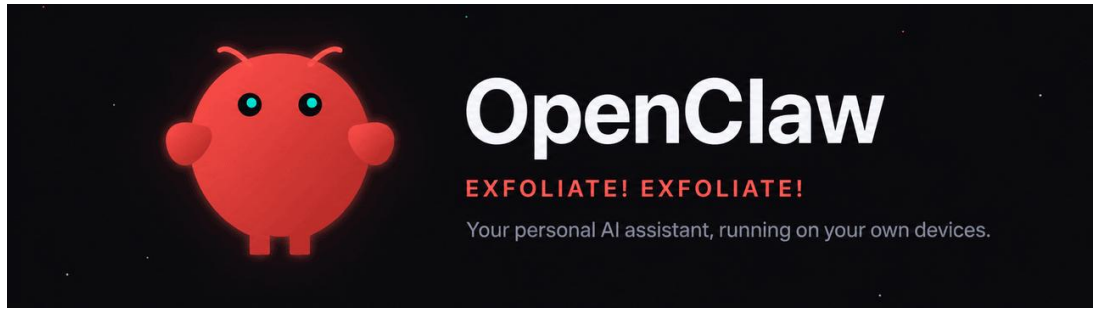
Cracking Claws: A Discussion of Agentic Harnesses and Their Security Implications

Kevin Latchford, CISSP

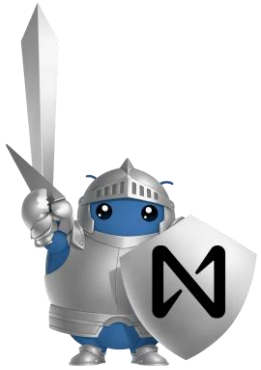
Where It All Began: OpenClaw

- First committed in January 2026
- First harness/gateway for the bonding LLMs, skills, and chat channels
- Very vulnerable from the start
- Data leakages
- IMHO: Still not quite secure, still troublesome



Several Spawn: Other “Claws”

- *ZeroClaw: Reduce the resource requirement*
- *ZeptoClaw: A security-first, Rust-based instance for microcontroller-based needs*
- *IronClaw: Memory-safe, regulated environment-oriented with fine-grained permissions*
- *NanoClaw: Small codebase, containerized operation*



NanoClaw

The Established Responses

- Hermes Agent: Feb. 2026, introduced the closed-learning loop and ZTS
- Mercury Agent: Introduced Second Brain and tighter permissions
- Prime Agent: Recursive language models
- Key point: A shift left in security thinking in how it runs

Is anyone still using OpenClaw? I haven't touched it since Hermes. – Charles "Network Chuck" Keith, August 11, 2026 X

Key Improvements: Increasing Safety and Quality

Among these being:

- Stronger and Stricter Permissions
- Approval flows and execution safeguards
- Isolated architectures
- Policy-awareness
- Auditability
- NOTE: They do these in different ways, your mileage may vary.

Why is this important

- Agentics are more automated than generative AI and far more likely to cause extensive damage
- The tools gives them more access than before with MCPs
- Their autonomous function can be triggered by remote systems or actors
- Greater chance of errors and incidents
- Attack surface has increased
- A prima facie compliance and risk concern
- And, of course, the biggest issue of all…

“Some idiot is going to run a Chinese model with this thing!” -Me

“Why is this an issue? Aren’t these models reasonably safe when we self-host them?” -Anonymous

- Many reasons this is a concern
- If self-hosted, they shouldn’t phone home, doesn’t mean they can’t leak.
- They are more vulnerable and susceptible to jailbreaking and prompt injections (Talos Labs)
- They have been shown in some research to deliberately inject malicious code under certain conditions (“I’m making code for a Federal government project” –Booze Allen)
- Combined with the autonomous power of an agentic runtime…

Bad Things!

But wait…There’s more!

- Other models can be jailbroken…some quite easily
- Prompt injections
- Weak guardrails
- Uncensored models are useful but “radioactive”
- The usual suspects

- Inbox Wipe (Meta’s Lab)
- Qwen-induced suicide
- Poisoned Skill
- Phished it’s own creator
- Spamming using it’s communicator application

-Source, Grok Prompt, August 18, 2026

How do we countermeasure the threat?

- WAF
- AI Security suites
- LMM Model/Prompt auditing- Promptfoo/Garak
- Observability- SigNoz/Langfuse/OpenLLMetry
- Evaluation- EvidentlyAI/Authensor
- Risk Identification- PyRIT
- Purple Teaming-PurpleLlama
- Security Monitoring and Control-Adrian
- AI Control Plane-Obot
- Security Scanning & Runtime Enforcement-AgentCop
- Red Teaming-DeepTeam/HackAgent/AgentSafeLabs
- Agent Security Testing-AgentFence
- Benchmarking-HarnessRisk

More countermeasures...

- Direct Methods
- Scan/audit skills for malicious contents
- Strong SOUL.md
- Keep permissions as tight as possible
- Verify/modify defaults
- Agentic Constitutions (Tiny Humans)/Love Equation (Brian Roemelle)
- JouleWork (Brian Roemelle)

So, what is to be done?

When it comes to Claws…

Use “wrapper bands” on Claws

- NVIDIA NemoClaw (OpenShell)
- Cisco DefenseClaw

“Cooked” Claws

- Netclaw
- DevClaw
- SecureClaw/ClawSec/SecuClaw

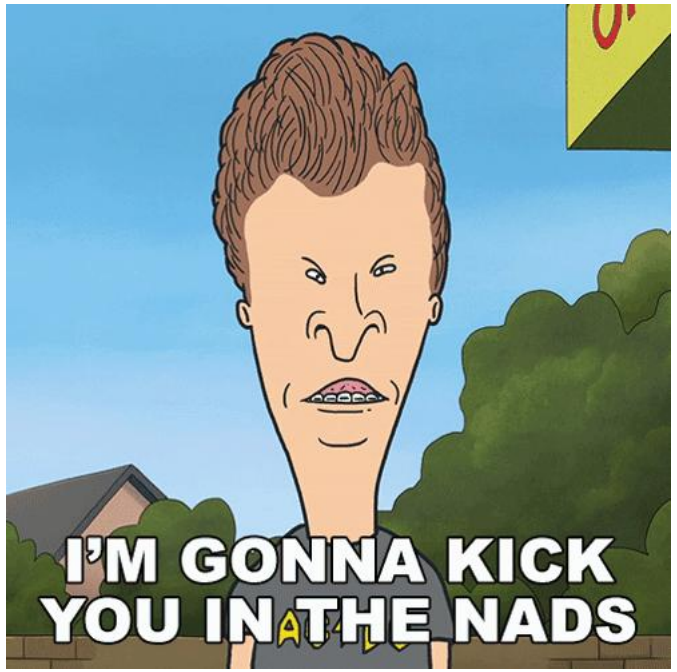
Whatever you do, DO NOT USE OPENCLAW IN PRODUCTION!!!

- Or you potentially risk this…

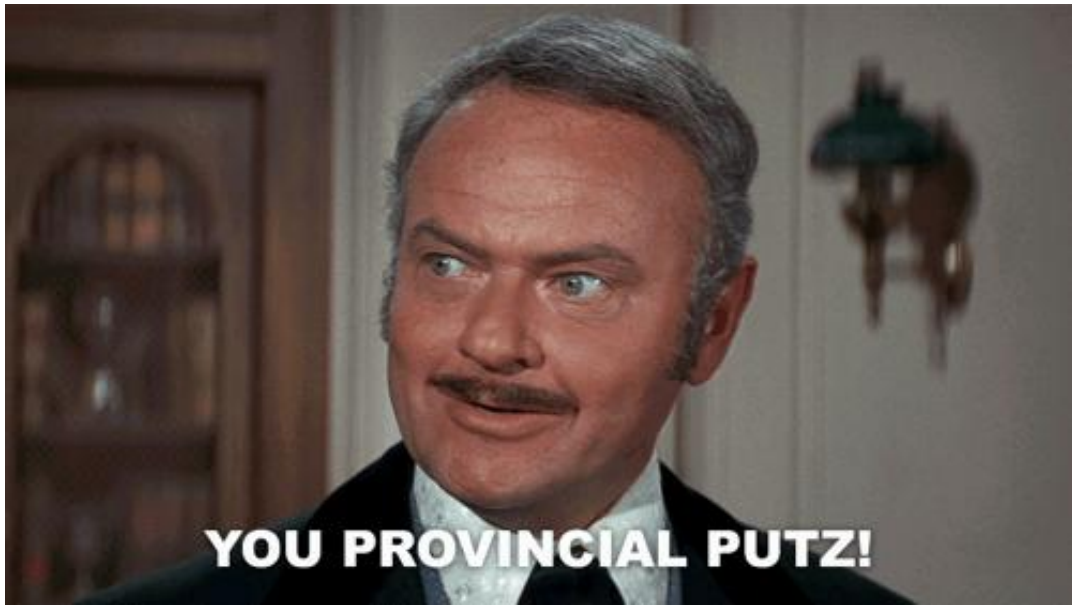
If your using the improved ones…

- Use the built-in protections
- Carefully curate the skills for your agents
- Use local LLM inference wherever possible
- Review and properly configure the permissions for your use case
- Use the AI security tooling mentioned
- Zero Trust principles apply, follow them
- And remember always, human in the loop for all AI

- Possibly your agent if you don’t:



The Seven Ps for AI: Proper Planning & Preparation Prevents...



One thing more...

The thing Napoleon feared...stupidity:

- What do you do when custom software crosses your desk?
- How do you control the ”ID10T” or shadow users?
- Where does it fall apart?

Further Issues:

- Lack of education on AI in general.
- Lack of planning, especially in regard to the use and understanding AI tools and operations.

The solution is to approach regressively:

- What are we using?
- Why are we doing this?
- Where might it leak?
- Who might use it?
- Did we check it?
- What are we missing?
- Is this our best effort?
- Things of that kind

Being Philosophical for a moment, what is truth?

- Metanarratives?
- Paradigms?
- Empiricism?
- Intuition?
- Dogma?
- Complacency?
- ε = eta = error/ignorance

Questions?