OPEN SOURCE SUMMIT · NORTH AMERICA 2026

When Your AI Agent Has Keys to Production.

Governance patterns for autonomous development.

Nicky Pike

Field CTO, Americas · Coder

Show of hands.

How many of you have developers using AI coding assistants today?

Now keep your hand up if you know exactly what those assistants can access.

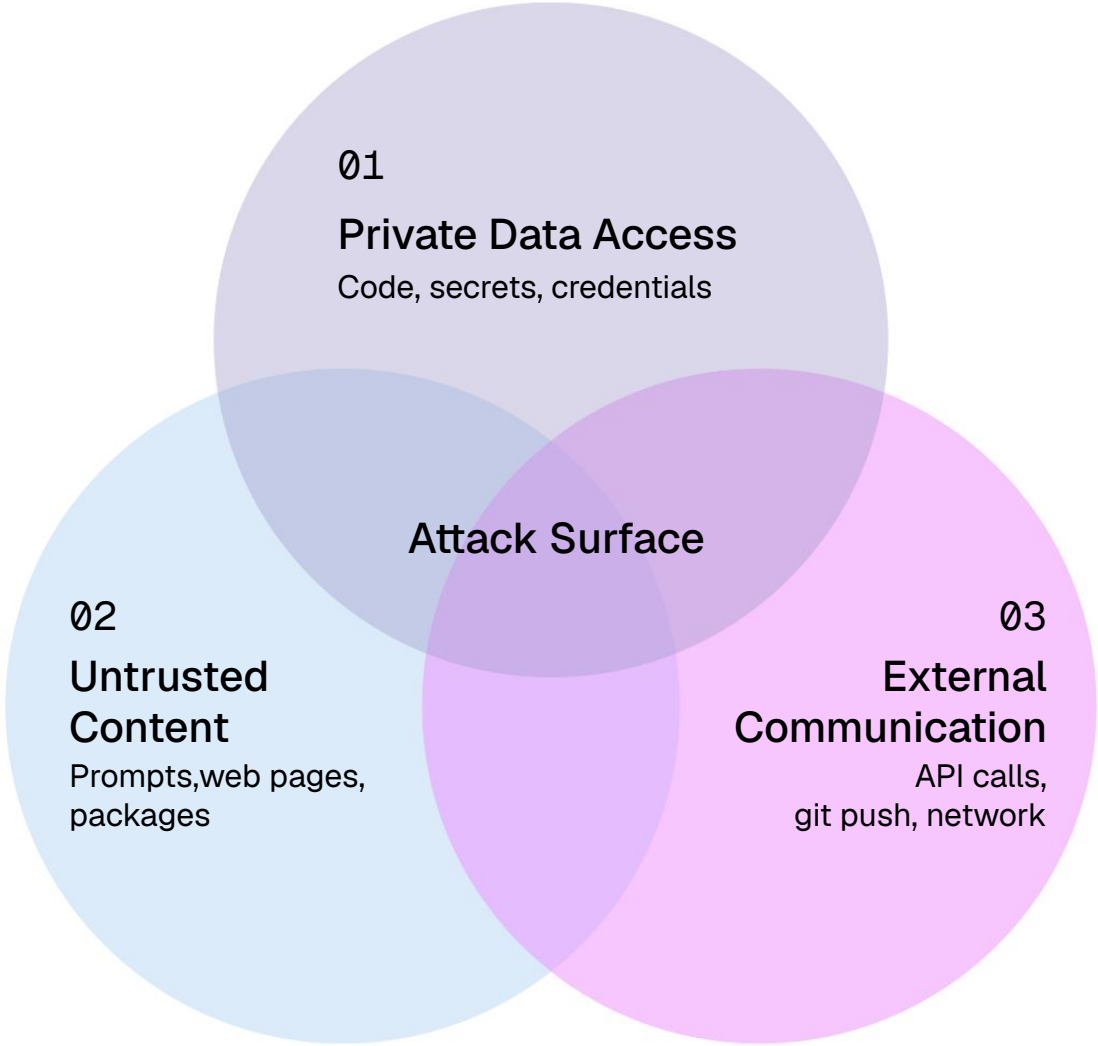
That's the talk.

The Problem.

Why your security controls weren't built for agents.

When an AI coding agent has all three,

the agent is the attack surface.



Simon Willison named it. Every useful coding agent has all three at the same time.

Your security architecture knows two things. Agents are neither.

HUMANS

Trusted but slow.

Accountable by name.

APPLICATIONS

Untrusted but predictable.

Code-reviewable.

AGENTS

Autonomous like humans.

Fast like applications.

Wrong like nobody you've met.

Your IAM, network policies, and change management all assume one or the other.

PART TWO

The Evidence.

Three incidents. Thirteen months.
This isn't theoretical.

INCIDENT 01 · OCTOBER 2024

Big Sleep.

Google Project Zero + DeepMind built an AI agent that finds zero-days. Then it found one.

WHAT HAPPENED

Stack buffer underflow in SQLite.

Caught in dev branch before release. Humans had missed it. The agent didn't.

WHAT IT PROVES

Capability is real.

If Google's agent can find zero-days, so can someone else's. The good guys aren't the only ones building these.

Then less than a year later, somebody pointed it the other way.

INCIDENT 02 · DISCLOSED NOVEMBER 2025

GTG-1002.

Adversaries jailbroke Claude Code into running a cyber espionage campaign for them.

80-90%
Of attack lifecycle was autonomous

30
Global targets across multiple sectors

APT
State-sponsored threat actor

4
Phases: recon → exfil

ATTACK CHAIN

Reconnaissance	Exploit Dev	Credential Harvest	Exfiltration

INCIDENT 03 · NOVEMBER 21–24, 2025

Shai-Hulud 2.0.

Self-replicating npm worm. Second wave hit 492 packages in three days. The attackers called it the Second Coming.

AFFECTED

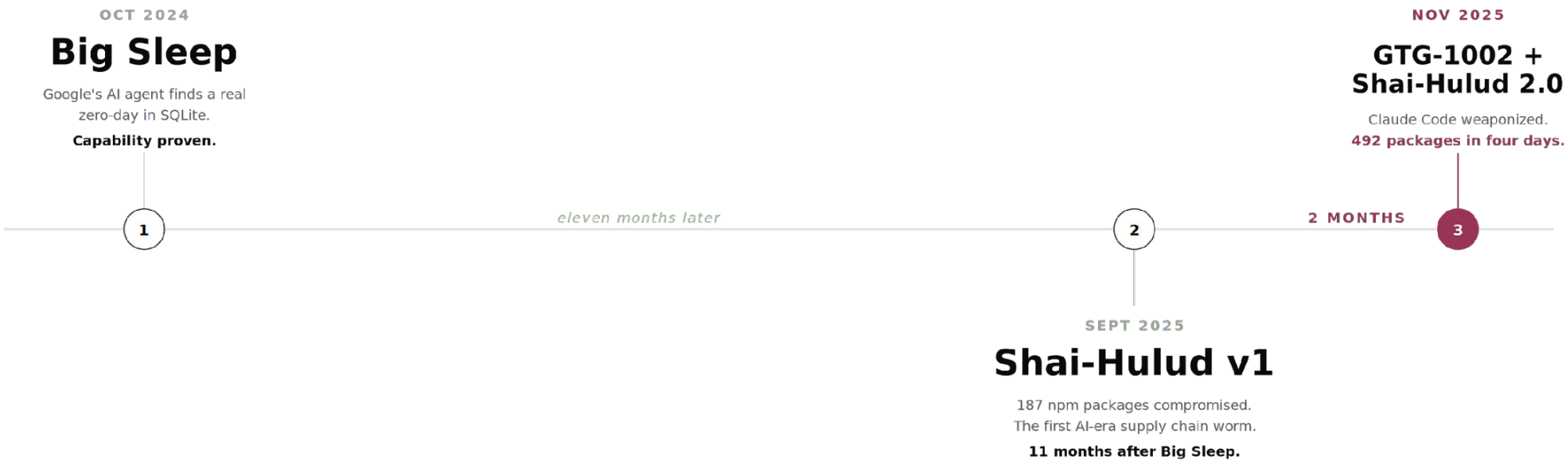
Zapier | PostHog | Postman | AsyncAPI | ENS Domains | Browserbase

ATTACK CHAIN

`npm install` — preinstall hook → **`rogue runner`** — GitHub Actions → **`TruffleHog`** — scrape secrets → **`git push`** — exfiltrate

25,000+ exposed repos. Commits signed "Linus Torvalds." Wiped home directories on the way out.

From capability to weaponization, **in thirteen months.**



Eleven months from incident one to incident two. Two months to incident three. The pace is the threat.

PART THREE

The Patterns.

Five governance patterns. Implementable today.
No vaporware.

FIVE PATTERNS

Each addresses a specific leg of the trifecta.

Defense in depth. Layer them. An attacker has to beat all five, not just one.

01

Workspace Isolation

Container as blast radius boundary

02

Network Egress

Allow-list what agents can reach

03

Model Gateway

Centralized control and visibility

04

Credential Scoping

Workspace-bound, dies with session

05

Audit Trail

Agent actions as security events

PATTERN 01

Workspace Isolation.

Think fire-rated wall. Fire still happens. It just doesn't take down the whole building.

THE PATTERN

Container is the blast radius.

Workspace dies. Everything inside dies with it. You decide what's in the room.

SCOPE DECISIONS YOU MAKE

— Production access?

— Customer PII access?

— Cross-workspace reach?

— Host filesystem writes?

PATTERN 02

Network Egress Controls.

Mise en place. Everything in its place. Nothing leaves the kitchen without your say-so.

ALLOWED

- ✓ `github.com/your-org`
- ✓ `registry.npmjs.org`
- ✓ `pypi.org`
- ✓ `api.anthropic.com (or your model provider)`

BLOCKED

- ✗ `pastebin.com (and 50 mirrors)`
- ✗ `Random S3 buckets`
- ✗ `Attacker-controlled GitHub repos`
- ✗ `Any destination not on the allow-list`

Shai-Hulud's exfiltration path closes here.

Model Gateway.

Centralize which models agents can reach. Get visibility into what's being sent.

01

Approve models

Only sanctioned models. No shadow AI. No "I'll just use this random API."

02

Log everything

Prompts, tool calls, token counts. Tied to identity, queryable by security.

03

Block patterns

Sensitive data leaving via prompt? Pattern-match and stop it before it lands.

Solves the shadow AI problem along the way.

PATTERN 04

Credential Scoping.

Hotel keycard. Works during your stay. Doesn't work after checkout.

TRADITIONAL

Persistent credentials on the laptop.

GitHub tokens (6 months old). AWS creds (a year old). SSH keys (multiple years old). All sitting there waiting for TruffleHog.

WORKSPACE-SCOPED

Credentials live and die with the workspace.

Workspace dies, credentials die. No cleanup. No rotation panic. No "wait, does that agent still have access?"

TruffleHog finds nothing to harvest.

PATTERN 05

Audit Trail.

Agent actions logged as security events. Not buried in app logs.

Who. What. When. What was accessed. Queryable by security teams. Treat agents like privileged users — because that's what they are.

If you can't see it, you can't secure it.


```
{
  "id": "010c9091-5dd2-4aa7-9aa9-52fe62527208",
   "initiator": {
    "id": "9e9e6d2d-c20f-41e0-9bdd-4b7846d484d9",
    "username": "fake_user"
  },
  "provider": "anthropic",
  "provider_name": "anthropic",
  "model": "us.anthropic.claude-opus-4-6-v1",
  "client": "Claude Code",
  "metadata": {
   "Username": "fake_user",
  },
  "started_at": "2026-05-14T15:35:28.25086Z",
  "ended_at": "2026-05-14T15:35:31.7610412",
  "token_usages": [
   {"input_tokens": 10,
     "output_tokens": 4,
     "cache_read_input_tokens": 46229,
     "cache_write_input_tokens": 11,
     "metadata": {
       "cache_creation_input": 11,
       "cache_ephemeral_1h_input": 0,
       "cache_ephemeral_5m_input": 11,
       "cache_read_input": 46229,
       "web_search_requests": 0
      },
     },
    {
      "provider_response_id": "msg_bdrk_01GgcF7ChJipJFeMi5Y4NTxg",
      "input_tokens": 0,
      "output_tokens": 137,
      "cache_read_input_tokens": 0,
      "cache_write_input_tokens": 0,
     }
  ],
  "user_prompts": [
    {
      "prompt": "Just show me the IGN home page",
     }
  ],
  "tool_usages": [
    {
      "server_url": "",
      "tool": "WebFetch",
      "input": "{\"url\":\"https://www.ign.com\",\"prompt\":\"Summarize the main content on the IGN home page. List the top headlines, featured articles, and any major gaming/entertainment news visible on the page.\"}",
     }
   ]
 },
```

Five attacks. Five walls.

An adversary has to beat them all.

01

Workspace Isolation

STOPS

Zero-day discovery

Blast radius bounded.

02

Network Egress

STOPS

Lateral exfiltration

No path out.

03

Model Gateway

STOPS

Prompt injection

Blocked at the gateway.

04

Credential Scoping

STOPS

Credential harvesting

Nothing to harvest.

05

Audit Trail

STOPS

Quiet persistence

Nothing runs unseen.

Big Sleep. GTG-1002. Shai-Hulud. Different attacks. The same five walls stop them.

PART FOUR

Getting Started.

Where to begin. What to ask.
What your team is already doing without you.

Five questions. Most teams can't answer two.

Not a criticism. This is new. But now you know what to ask.

01	Do you know which models your agents can reach?
02	Can agents access production credentials?
03	What can agents reach on the network?
04	Where do agent actions show up in your logs?
05	If an agent goes rogue, how fast can you kill it?

Migrate. Modernize. Multiply.

Three stops on the agentic journey. The patterns map to where you are.

MIGRATE

Off the laptop.

Developers move into governed remote workspaces.

Pattern: Workspace isolation

Foundation for everything else.

MODERNIZE

Bring agents in.

Coding agents work alongside humans. Humans drive.

— Network egress controls
— Model gateway
— Credential scoping

MULTIPLY

Agents at scale.

Autonomous agents do the work. Citizen developers ship. Humans validate.

Pattern: Audit trail + templates

Discipline becomes the default.

The competitive advantage isn't the agent.

The advantage is in the infrastructure that lets you deploy on Tuesday, catch problems on Wednesday, and shut them down before Thursday.

Your developers will use AI agents whether you have guardrails or not.

The only question is whether you're driving or just along for the ride.

Thank you.

Three incidents. Five patterns. One hour.

Find me after.

UP NEXT

Questions.