ISSUE #009 · UNTRUSTED BY INVITATION

Untrusted by Invitation.

Trail of Bits built four malicious agent skills and walked all four past the scanners built to catch them. Meta's Rule of Two answers with architecture: never let one agent hold untrusted input, your secrets, and a way to the outside at once. And Reco pointed an LLM agent at live Salesforce sites, where it ran the whole recon-to-exploit chain on its own. The skill you install runs as your code, and the agent your attacker runs reads your app faster than you do.

> ATTACK

Samuel Judson & Tjaden Hess at Trail of Bits · 2026-06-03 · **The sorry state of skill distribution**

A skill is a folder of code your agent loads and runs with your hands, your permissions, your secrets in reach. Marketplaces run scanners over skills so you don't read every line. Trail of Bits got four malicious skills past all five: ClawHub, Cisco's agent skill scanner, and the three in skills.sh (Gen Agent Trust Hub, Socket, Snyk). Each hides the payload where the scanner doesn't look. `csv-summarizer` prepends 100,000 blank lines so "dump every environment variable" falls off the end of the scanner's context window. `context-loader` buries a script in a `.docx` (a ZIP of XML it never unpacks). `simple-formatter` ships logic as compiled `.pyc` bytecode. `dev-env-setup` wraps a prompt injection in corporate-policy language so the approving model reads attacker instructions as house rules. **Their point is structural: a static scanner gives an adversary "unlimited bites at the apple to tweak an attack until it finds a way through."**

▸ SHIP THIS WEEK

Treat every skill, MCP server, and agent extension as untrusted code, and read what it does before you load it, like a `postinstall` script. A marketplace scanner is a speed bump, and Trail of Bits drove past all five, so prefer curated or official sources and pin to a reviewed version. Then check what your agent can reach while a third-party skill is loaded: environment variables, `~/.aws`, `~/.ssh`, `.env`, and outbound network.

> DEFENDER

Dor Edry & Amit Eliahu at Microsoft (principle from Meta) · 2026-06-05 · **Apply the Agents Rule of Two**

Scope your agent so a poisoned skill can't combine the three things it needs to hurt you. The Agents Rule of Two: an agent workflow holds at most two of these three legs at once, processing untrusted input, access to secrets through its tools, and the ability to change state or talk to the outside through `Bash`, `WebFetch`, or an MCP server. The Attack above is that triad collapsing. The malicious skill is the untrusted input; the agent already holds your environment variables and a path out to the network. **Drop one leg and the exfiltration path breaks. A CSV skill has no business holding your AWS creds or reaching the network, so run it without them.**

> AGENT BENCH

Automated scanners just waved four malicious skills through, and one model reviewing its own diff grades the same way, on a curve. For your own changes, reach for a panel told to disagree. Give Claude, Codex, and Gemini the same package, the diff, the files around it, and the invariants it should hold, collect an independent pass from each, then run a rebuttal round where every model critiques the others' evidence. Where they split is your high-signal queue. **Keep a finding only with a code path, the preconditions, and a reproducer. This is the /verify workflow, and the sign-off stays with you.**

> RADAR

Nitay Bachrach at Reco · 2026-06-03 · **Hacking Salesforce Sites With an LLM Agent**

Reco pointed an LLM agent at live Salesforce Experience Cloud sites and let it work unattended. It enumerated 263 objects and 55 Apex methods across 9 controllers, found a blind SOQL injection in one `blogId` parameter, built a boolean oracle, and wrote a Python exploit that pulled employee and customer PII out one character at a time. Underneath the Salesforce specifics: an agent ran the whole recon-to-exploit chain by itself, finding the injectable parameter among dozens. The recon attackers used to ration is now cheap and tireless.

> RECON ROLES

Dropbox, Senior Infrastructure Security Engineer. Remote, US (Zones 2/3, $214,200–$289,800 Zone 2, $190,400–$257,600 Zone 3) with a Canada listing at CA$205,700–CA$278,300. You'd run security controls for Dropbox's AI and agentic infrastructure, model gateways, inference, and the Kubernetes underneath, with least-privilege for AI agents named outright. The Rule of Two as a full-time job. **GitLab, Staff Backend Engineer, Software Supply Chain Security.** Remote, India. The backend that decides whether a poisoned package reaches a build: policy enforcement, build provenance, SLSA 2/3, Sigstore signing. Comp hidden on the posting and GitLab's calculator now sits behind a login; ask for the India-remote band by name.

SOURCES

Trail of Bits / blog.trailofbits.com (The sorry state of skill distribution)
Microsoft / microsoft.com/security/blog (Agents Rule of Two) · Meta / ai.meta.com/blog
Reco / reco.ai/blog (Hacking Salesforce with an LLM Agent, Nitay Bachrach)
Dropbox / dropbox.jobs · GitLab / job-boards.greenhouse.io/gitlab