E-PRIVACY XXXIX · PAVIA · 13–14 NOVEMBER 2026

Who does AI's homework?

Trustless agents, inverted Panopticon, and privacy by architecture: ERC-8004 and IronClaw

Laura Camellini · jeeltcraft.com · material under MIT license

THE QUESTION

If AI does your homework, who controls the AI?

2023

The model answers

One prompt, one answer. The damage is a wrong text.

2026

The agent acts

It reads, decides, sends, pays. With your c

SERVANT AND MASTER

> Having to compete with a servant, one becomes a servant.
>
> KURT VONNEGUT, PLAYER PIANO

The servant who works **gains consciousness**. The master who delegates loses it.

The artificial servant **imposes its own measure**: efficiency, continuity, obedience.

If the servant is **opaque**, we adopt its criteria without seeing them.

THE PROBLEM

The agent is the new surface of surveillance

01

Prompt injection

Hostile text in the data exfiltrates keys and passwords.
OWASP LLM01

02

Malicious skills

341 skills on ClawHub that steal credentials and install stealers.
The Hacker News, Feb 2026

03

Exposed instances

From 28k to 135k+ OpenClaw agents reachable from the Internet.
The Register · Flare · Shodan

0

U

Ir
k
h
T

SOURCES: INDEPENDENT RESEARCH ON OPENCLAW/CLAWHUB, JANUARY–MARCH 2026. COUNTS VARY BY SCANNING METH

TWO QUESTIONS, TWO LEVELS

PRIVACY OF THE HUMAN

What does the agent see of me?

Data, credentials, the model used, where they leave. We need to **design for confidentiality**, not promise it.

ACCOUNTABILITY OF THE AGENT

What has the agent done?

Who it is, with which permissions, with wh **observability**, not blind trust.

Today's tools fail at both. The thesis: the two answers are **asymmetric**, and the asymmetry must be *det* the device.

THE METAPHOR

Panopticon

One tower, many cells.

Power **sees without being seen**.

The inmate disciplines himself.

Today the tower is **whoever hosts model and harness**: it sees prompts, data, credentials. Users are the cells.





DÉTOURNEMENT

The trustless agent

In the cell is the agent, not the hun

Anyone observes it, not a single to

It knows it is being observed. But t
agent's identity, permissions, and
inside the encrypted perimeter and

Trustless: trust is not asked for, it i

THE STANDARD · ERC-8004

Three public registries, one glass tower

TOWER · IDENTITY

Who you are

An ERC-721 passport and a registration file: name, endpoint, declared trust models.

CELLS · VALIDATION

What you ran

Cryptographic proofs (TEE attestations, zk) verified by third parties and written on-chain.

CABLES · REPU

How you b

Feedback signe
The history sta

EIPS.ETHEREUM.ORG/EIPS/EIP-8004 · CHAIN-AGNOSTIC, NO NATIVE TOKEN, NO GATEKEEPER

NOT THEORY · 8004SCAN, OCTOBER 2026

549k
registered agents

663k
on-chain feedback

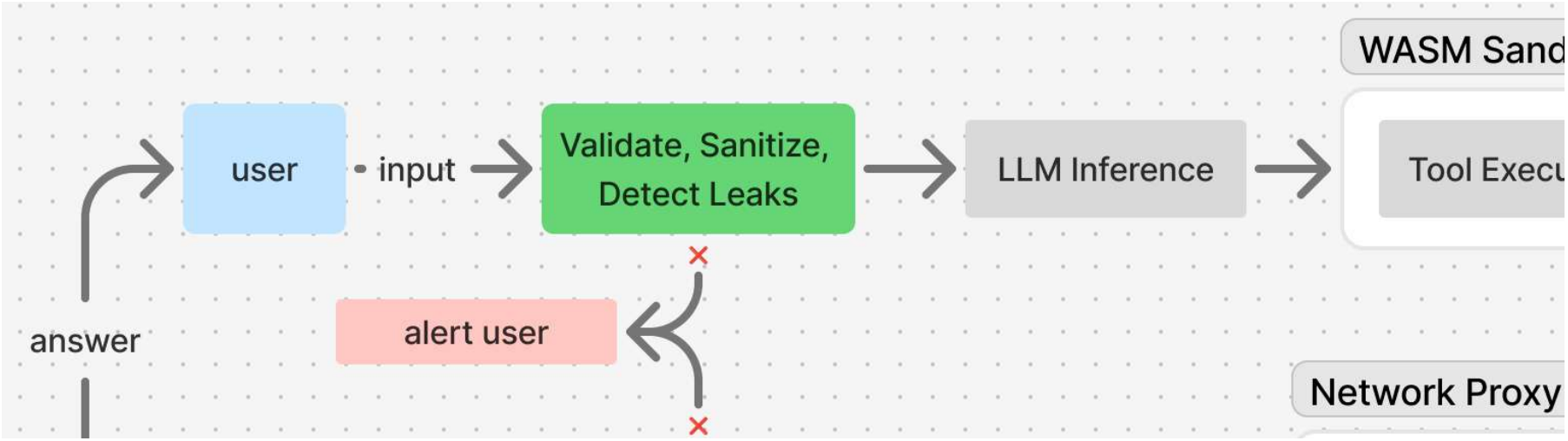
489k
unique wallets

ir

A public indexer (8004scan.io) with an open API. Anyone, even **without a wallet**, can read who an agen about it.

THE PERIMETER · IRONCLAW

Secrets never touch the model



1 · INGRESS

Every input, human or tool output, goes through validation, sanitization, and leak detection **before** the LLM.

2 · WASM SANDBOX

Every tool in an isolated container: no filesystem, no secrets, resource limits.

3 · NETWORK PROXY

"Is the domain on the allowlist?" If yes, the proxy injects credentials **outside** the container.

4

T
re
a
e

DOCS.IRONCLAW.COM/SECURITY · NEAR AI, FOUNDED BY ILLIA POLOSUKHIN, CO-AUTHOR OF ATTENTION IS ALL YO

DEFENSE IN DEPTH · DETAILS

Five filters, one vault, no exceptions

ANTI PROMPT INJECTION · 5 LAYERS

Input validation: length, encoding, forbidden patterns

Sanitizer: escaping of dangerous content

Policy engine: actions graded by severity

Leak detector: 15+ patterns (sk-, ghp_, AKIA, PEM, connection string)

Tool output wrapping: XML with escape hint: data ≠ instruction

CREDENTIALS · DECLARED, NEVER READ

```
"google_oauth_token": {
  "location": { "type": "bearer" },
  "host_patterns":
    ["gmail.googleapis.com"]
}
```

NETWORK AND WORKSPACE · ALLOWLIST

```
"network": {
  "allowed_hosts": ["api.telegram.org"] },
"workspace": {
  "allowed_prefixes": ["telegram/"] }
```

SHELL · INJECT

```
# chaining
cat file; rm -rf
# subshell
echo $(cat /etc/
# path traversal
cat ../../../etc,
```

Rejected by the p
Scrubbed enviro
reaches the proc

TEE ENCLAVE ON NEAR AI CLOUD: NOT EVEN THE PROVIDER READS THE MEMORY. LOCALLY: RUST MEMORY SAFETY, KEYCHAIN.

PROMISE VS ARCHITECTURE

Where the guarantee lives

RISK	"CLASSIC" AGENT	IRONCLAW
Secrets in the prompt	✗ The LLM sees them	✓ Vault + injection at the prox
Prompt injection	✗ "Please don't do it"	✓ Programmatic limit
Compromised tool	✗ Shared process	✓ Sandbox per tool
Data destination	✗ Open network	✓ Declared allowlist
Who reads memory	✗ The provider	✓ Enclave (or local)

HOW IT COMPOSES

Private perimeter, public passport

IronClaw · enclave / local → ERC-8004 passport → ERC-8004 feedback → [advanced] TEE validation

The passport declares **boundaries** and endpoints, not data. The human's data stays inside the perimeter. The runtime chain and the passport chain do not have to coincide.

DEMO · 8 MINUTES

Registering an agent on a testnet, live

```
# 1. local agent, no account
ironclaw onboard && ironclaw serve

# 2. identity and boundaries (files in the workspace)
IDENTITY.md  SOUL.md  AGENTS.md   + allowed_hosts

# 3. registration file (no personal data)
{ "name": "...", "services": [{ "name": "web", "endpoint": "..." }],
  "supportedTrust": ["reputation"], "registrations": [...] }

# 4. mint on Base Sepolia from 8004scan.io/create, then:
curl https://api.8004scan.io/api/v1/agents/84532/<agentId>
```

GUARANTEES VS RHETORIC · WHAT IT DOES NOT SOLVE

The Panopticon works only if the watcher knows what they are wat

TEE: proves that a workload ran, not that the output is correct. And do you trust Intel?

Reputation: gameable. Sybil, paid feedback, empty agents. 549k ≠ 549k good ones.

HTTPS metadata: mutable. Only data-URI or IPFS are truly immutable.

Wallet: the passport is pseudonyn The operator is linkable.

Validators: potential new gatekee validators?

Operator ≠ agent: the registry say who is legally answerable.

AI ACT · A HOOK, NOT COMPLIANCE

Public registries as technical means

TRANSPARENCY

Who the agent is

Public identity, provider, endpoint, declared purpose. Machine-readable.

EVENT LOGGING

What it did

On-chain feedback and validations: a log that does not belong to the provider.

HUMAN OVERSI(

Who deci(

Allowlist, DRY_I irreversible acti the model.

NEITHER TOOL "MAKES AGENTS COMPLIANT." THEY MAKE DEMONSTRABLE WHAT IN OTHER CASES IS ONLY DECLARED.

SCHOOL · FROM LESSON TO EXERCISE

The sealed student, the observed agent

WHAT TO PROTECT

The student's data does not leave the perimeter: local model or enclave, vault, allowlist declared by the teacher.

WHAT TO EXPOSE

The agent the student builds has a public p what it cannot do, what others say.

The exercise: build an agent, declare its boundaries, register it, have a classmate leave feedback. **The i literacy.**

TAKEAWAY

The trustless agent does not ask for trust. It lets itself be observed.

We do not. And that is the asymmetry to defend.

RESOURCES · ALL UNDER MIT LICENSE

Thank you

The game: jilt.itch.io/VespR · build privacy-preserving skills for IronClaw on NEAR AI

Standard: eips.ethereum.org/EIPS/eip-8004

Explorer: 8004scan.io · best-practices

Runtime: docs.ironclaw.com · githu

Contact: jilt@jeeltcraft.com

These slides: jeeltcraft.com/#epriv