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

Chi fa i compiti dell'IA?

Agenti trustless, Panopticon rovesciato e privacy by architecture: ERC-8004 e IronClaw

Laura Camellini · jeeltcraft.com · materiale sotto licenza MIT

LA DOMANDA

Se l'IA fa i tuoi compiti, chi controlla l'IA?

2023

Il modello risponde

Un prompt, una risposta. Il danno è un testo sbagliato.

2026

L'agente agisce

Legge, decide, invia, paga. Con le vostre c

SERVO E PADRONE

> Dovendo competere con un servo, si diventa servi.
>
> KURT VONNEGUT, PIANO MECCANICO

Il servo che lavora **conquista la coscienza**. Il padrone che delega la perde.

Il servo artificiale **impone la propria misura**: efficienza, continuità, obbedienza.

Se il servo è **opaco**, adottiamo i suoi criteri senza vederli.

IL PROBLEMA

L’agente è la nuova superficie di sorveglianza

01

Prompt injection

Un testo ostile nei dati fa esfiltrare chiavi e password.
OWASP LLM01

02

Skill malevole

341 skill su ClawHub che rubano credenziali e installano stealer.
The Hacker News, feb 2026

03

Istanze esposte

Da 28k a 135k+ agenti OpenClaw raggiungibili da Internet.
The Register · Flare · Shodan

0

F
C
D
A
c
I

FONTI: RICERCHE INDIPENDENTI SU OPENCLAW/CLAWHUB, GENNAIO–MARZO 2026. I CONTEGGI VARIANO PER METODO

DUE DOMANDE, DUE LIVELLI

PRIVACY DELL'UMANO

Cosa vede l’agente di me?

Dati, credenziali, modello usato, dove escono. Serve **progettare per la confidenzialità**, non prometterla.

RESPONSABILITÀ DELL'AGENTE

Cosa ha fatto l’agente?

Chi è, con quali permessi, con quale storia **pubblica**, non fiducia cieca.

Gli strumenti attuali falliscono su entrambe. La tesi: le due risposte sono **asimmetriche**, e l'asimmetria v contro il dispositivo.

LA METAFORA

Panopticon

Una torre, molte celle.

Il potere **vede senza essere visto**.

Il detenuto si disciplina da solo.

Oggi la torre è **chi ospita modello e harness**: vede prompt, dati, credenziali. Gli utenti sono le celle.





DÉTOURNEMENT

L’agente trustless

Nella cella c'è **l'agente**, non l'umar

Lo osserva **chiunque**, non una sing

Sa di essere osservato. Ma nel reg
permessi e feedback dell'agente:
perimetro cifrato e **non vi entrano**.

Trustless: la fiducia **non si chiede,**

LO STANDARD · ERC-8004

Tre registri pubblici, una torre di vetro

TORRE · IDENTITY

Chi sei

Un passaporto ERC-721 e un file di registrazione: nome, endpoint, modelli di fiducia dichiarati.

CELLE · VALIDATION

Cosa hai eseguito

Prove crittografiche (attestazioni TEE, zk) verificate da terzi e scritte on-chain.

CAVI · REPUT

Come ti se

Feedback firma
storia resta legg

EIPS.ETHEREUM.ORG/EIPS/EIP-8004 · CHAIN-AGNOSTIC, NESSUN TOKEN NATIVO, NESSUN GATEKEEPER

NON È TEORIA · 8004SCAN, OTTOBRE 2026

549k
agenti registrati

663k
feedback on-chain

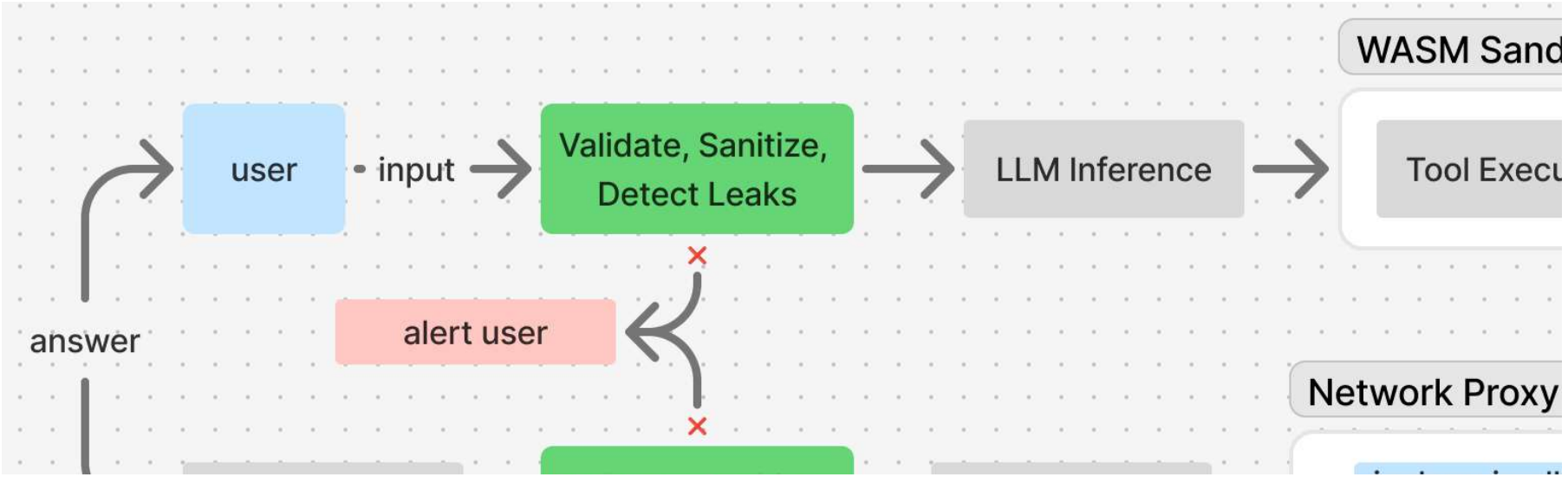
489k
wallet unici

Bl

Un indexer pubblico (8004scan.io) con API aperta. Chiunque, anche **senza wallet**, può leggere chi è un
lui.

IL PERIMETRO · IRONCLAW

I segreti non toccano mai il modello



1 · INGRESSO

Ogni input, umano o output di tool, passa da validazione, sanitizzazione e leak detector **prima** dell'LLM.

2 · SANDBOX WASM

Ogni tool in un container isolato: niente filesystem, niente segreti, limiti di risorse.

3 · PROXY DI RETE

"Il dominio è in allowlist?" Se sì, il proxy inietta le credenziali **fuori** dal container.

4

L
e
v
d

DOCS.IRONCLAW.COM/SECURITY · NEAR AI, FONDATA DA ILLIA POLOSUKHIN, COAUTORE DI ATTENTION IS ALL YOU

DIFESA IN PROFONDITÀ · DETTAGLI

Cinque filtri, un vault, nessuna eccezione

ANTI PROMPT INJECTION · 5 LIVELLI

Validazione input: lunghezza, encoding, pattern vietati

Sanitizer: escape dei contenuti pericolosi

Policy engine: azioni graduate per severità

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

Wrapping output tool: XML con escape hint: dato ≠ istruzione

CREDENZIALI · DICHIARATE, MAI LETTE

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

RETE E 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,
```

Rifiutati dal parse
Ambiente scrubb
segreta passa al

ENCLAVE TEE SU NEAR AI CLOUD: NEMMENO IL PROVIDER LEGGE LA MEMORIA. IN LOCALE: RUST MEMORY-SAFE, CH
DI SISTEMA.

PROMESSA VS ARCHITETTURA

Dove sta la garanzia

RISCHIO	AGENTE "CLASSICO"	IRONCLAW
Segreti nel prompt	✗ L'LLM li vede	✓ Vault + iniezione al pro
Prompt injection	✗ "Per favore non farlo"	✓ Limite programmatico
Tool compromesso	✗ Processo condiviso	✓ Sandbox per tool
Destinazione dati	✗ Rete libera	✓ Allowlist dichiarata
Chi legge la memoria	✗ Il provider	✓ Enclave (o locale)

COME SI COMPONE

Perimetro privato, passaporto pubblico

IronClaw · enclave / locale → Passaporto ERC-8004 → Feedback ERC-8004 → [avanzato] Validazione

Il passaporto dichiara **confini** ed endpoint, non dati. I dati dell'umano restano nel perimetro. La blockchain del runtime e quella del passaporto non devono coincidere.

DEMO · 8 MINUTI

Registrare un agente su una testnet, dal vivo

```
# 1. agente in locale, nessun account
ironclaw onboard && ironclaw serve

# 2. identità e confini (file nel workspace)
IDENTITY.md  SOUL.md  AGENTS.md   + allowed_hosts

# 3. file di registrazione (nessun dato personale)
{ "name": "...", "services": [{ "name": "web", "endpoint": "..." }],
  "supportedTrust": ["reputation"], "registrations": [...] }

# 4. mint su Base Sepolia da 8004scan.io/create, poi:
curl https://api.8004scan.io/api/v1/agents/84532/<agentId>
```

GARANZIE VS RETORICA · COSA NON RISOLVE

Il Panopticon funziona solo se chi guarda sa cosa guarda

TEE: prova che un workload è girato, non che l'output è corretto. E vi fidate di Intel?

Reputazione: gamabile. Sybil, feedback a pagamento, agenti vuoti. 549k ≠ 549k buoni.

Metadati HTTPS: mutabili. Solo data-URI o IPFS sono davvero immutabili.

Wallet: il passaporto è pseudonim L'operatore è collegabile.

Validatori: nuovi gatekeeper poter validatori?

Operatore ≠ agente: il registro dic risponde legalmente.

AI ACT · UN AGGANCIO, NON UNA CONFORMITÀ

Registri pubblici come mezzi tecnici

TRASPARENZA

Chi è l'agente

Identità pubblica, fornitore, endpoint, scopo dichiarato. Leggibile dalla macchina.

REGISTRAZIONE EVENTI

Cosa ha fatto

Feedback e validazioni on-chain: un log che non appartiene al fornitore.

SORVEGLIANZA

Chi decid

Allowlist, DRY_I azioni irreversib dal modello.

NESSUNO DEI DUE STRUMENTI "RENDE CONFORMI" GLI AGENTI. RENDONO DIMOSTRABILE CIÒ CHE IN ALTRI CASI S

SCUOLA · DALLA LEZIONE ALL'ESERCIZIO

Lo studente sigillato, l’agente osservato

COSA PROTEGGERE

I dati dello studente non lasciano il perimetro: modello locale o enclave, vault, allowlist dichiarata dal docente.

COSA ESPORRE

L'agente che lo studente costruisce ha un può fare, cosa non può, cosa dicono gli alt

L'esercizio: costruire un agente, dichiararne i confini, registrarlo, farsi lasciare un feedback da un comp **rovesciato come alfabetizzazione.**

TAKEAWAY

L’agente trustless non chiede fiducia. Si lascia osservare.

Noi no. Ed è questa l’asimmetria da difendere.

RISORSE · TUTTO SOTTO LICENZA MIT

Grazie

Il gioco: jilt.itch.io/VespR · costruisci skill privacy-preserving per IronClaw su NEAR AI

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

Explorer: 8004scan.io · best-practices

Runtime: docs.ironclaw.com · githu

Contatto: jilt@jeeltcraft.com

Queste slide: jeeltcraft.com/#epriv