OpenA2A

Securing AI agent infrastructure.

Identity, attestation, and behavioral trust at scale.

Abdel Fane

Founder and CEO, OpenA2A · WeAreDevelopers World Congress North America · 24 September 2026, Stage 5

WHO AM I

Abdel Fane.

20

years in technology

Founder and CEO, OpenA2A · Executive Director, CSNP

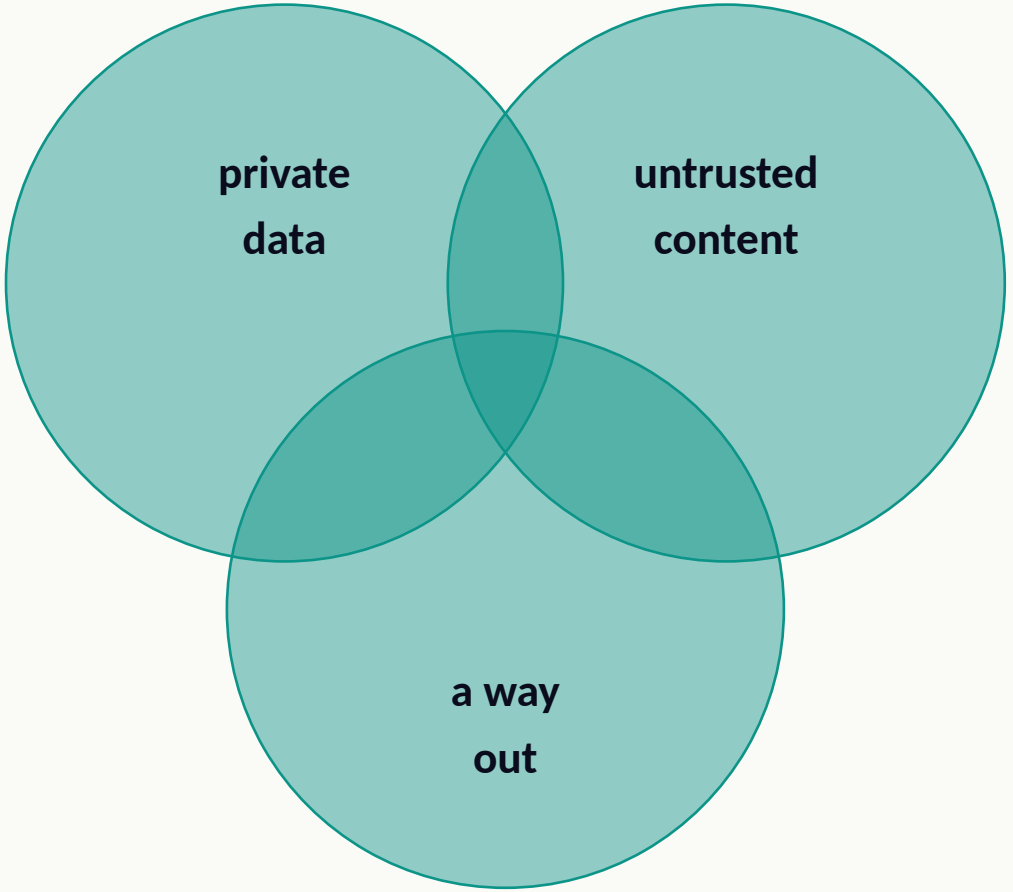
Booz Allen Hamilton and Protiviti. Allstate and Grail.

The open work

- A2A identity and trust contributor, PR #1496 (in flight)
- OpenTelemetry SemConv contributor, Issue #180
- Vulnerabilities fixed in OpenClaw after coordinated disclosure
- HackMyAgent core scanner contributed to OpenClaw (PR #9806)

A FRAME YOU ALREADY KNOW

The lethal trifecta.



Simon Willison, June 2025

Your agent has all three.

- It reads your database
- It reads documents strangers wrote
- and it can make an HTTP request

The model did nothing wrong.

a reply on X,
in Morse code

Grok decodes it,
faithfully

Bankrbot treats the
output as a command

3 billion DRB tokens

about $150,000 to $174,000 at the time

No identity check and no capability check between the model and the money.

SlowMist, CryptoSlate, and Bankr statement, May 2026. About 80% later returned.

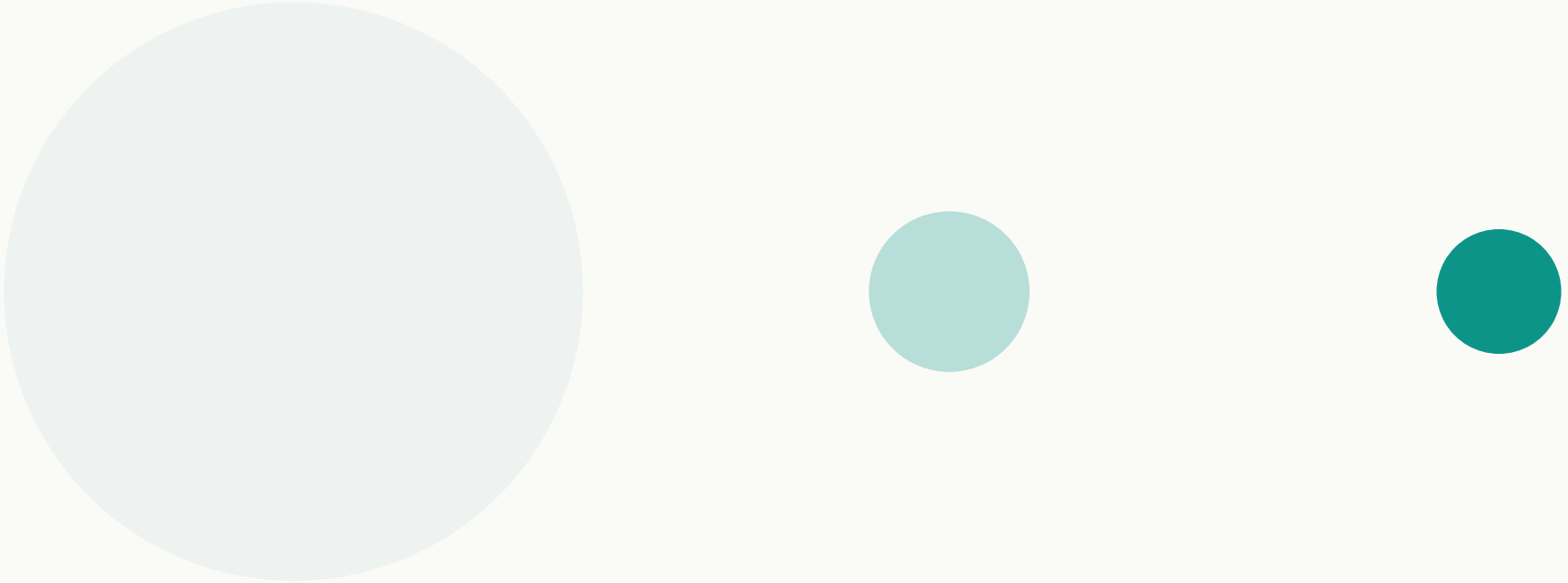
Stop securing the model.

Secure the tool call.

System Prompts are NOT security Guardrails, they’re suggestions

THE ARGUMENT, MEASURED

The tool call is where the signal is.



422,010

MCP connection events

35

MCP resource reads

9

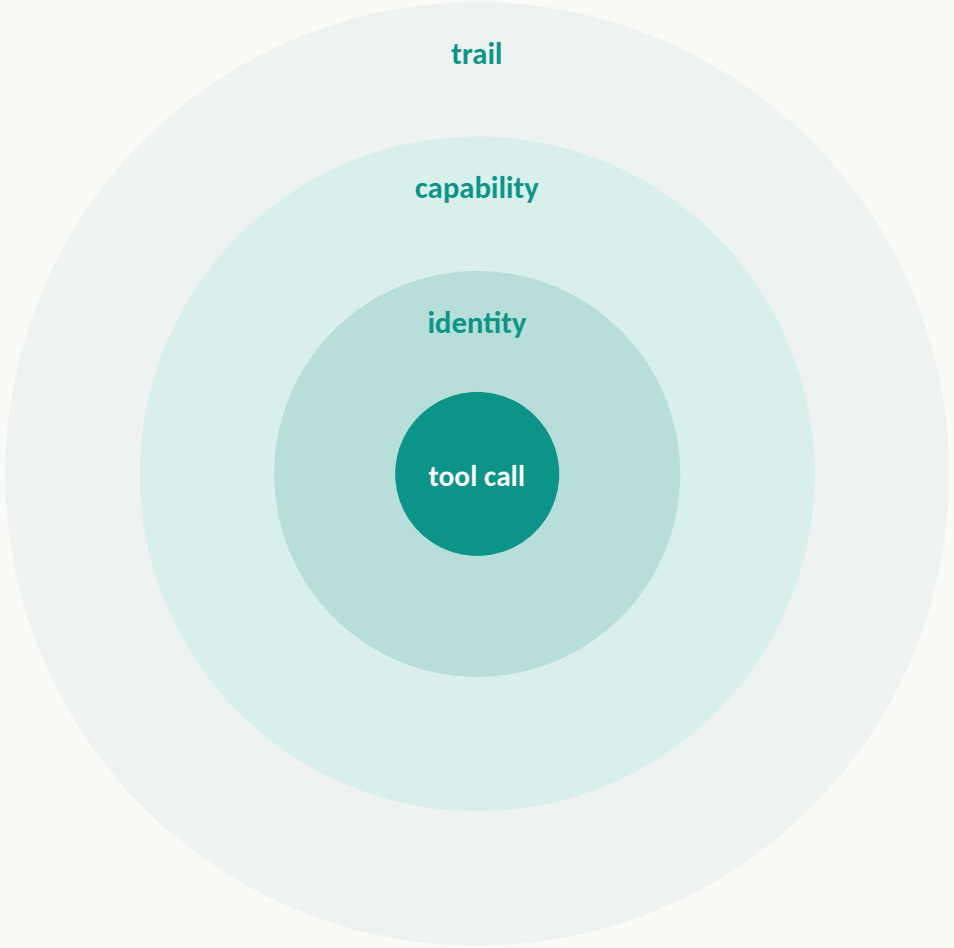
MCP tool calls

log scale

OpenA2A Behavioral Threat Report, Issue 5. 30 days to Sep 17, 2026.

THE ANSWER

Three rings around the tool call.



identity

A keypair the host generates. Not a string three services share.

capability

What this agent may do. Everything else is wall, not a rule.

trail

Every decision signed, allowed and refused alike.

A keypair, not a shared secret.

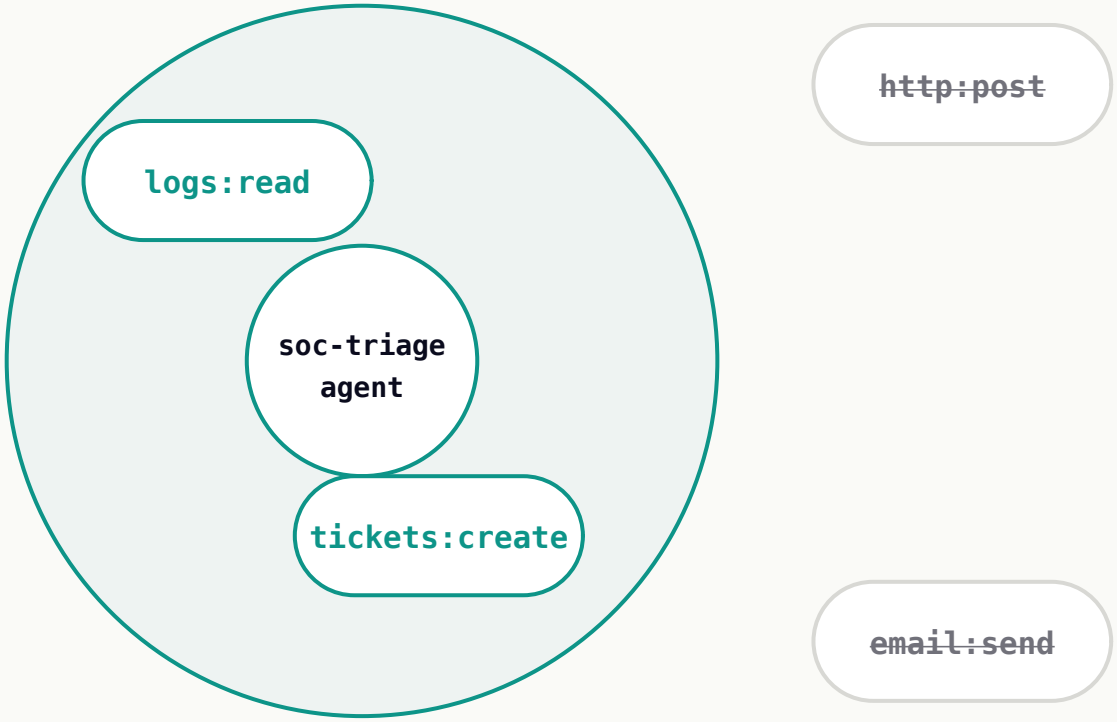
```
$ pip install aim-sdk


from aim_sdk import secure
agent = secure("soc-triage-agent")
```

- **Ed25519 keypair** — per agent, private half never sent
- **Named owner** — the logged in account
- **Pending** — until an admin verifies it

ATX credential with hybrid Ed25519 and ML-DSA-65 signatures. It survives the post quantum transition.

Everything outside the circle is wall.



Granted. proceed

logs:read and tickets:create, because that is the agent's job.

Never granted. blocked

Not blocked. Not detected. Nothing to catch.

Grok's wallet got its transfer rights from an NFT somebody sent it.

Know when your tools change under you.

14

MCP servers on one laptop
0 with a verified identity

1,997

reachable on the public internet, up from 1,775

12 → 14

tools on one server, between two runs

243,615

exposed AI services, up 10.1% on the month

```
postgres-mcp   attestation 46419f40… recorded   14 tools attested   drift: +tool_13 +tool_14 (severity low)
```

Supply Chain Security

OpenA2A Registry sweep, Sep 1, 2026. Tool line from a recorded run.

The trace shows the call, not the permission.


```
{"timestamp":"2026-09-24T17:41:02+00:00","category":"AUTHZ","event_type":"CAPABILITY_GRANTED",
 "agent_name":"soc-triage-agent","action":"logs:read","resource":"siem://alerts","result":"GRANTED"}

{"timestamp":"2026-09-24T17:41:09+00:00","category":"AUTHZ","event_type":"CAPABILITY_DENIED",
 "agent_name":"soc-triage-agent","action":"http:post","resource":"https://…","result":"DENIED",
 "details":{"http_status":403,"denial_reason":"capability not registered"}}
```


who what whether **proof**

A denial that leaves no evidence is indistinguishable from an attack that never happened.

LIVE DEMO

Same code. One has an identity.

01

Register

02

Grant

03

Run the job

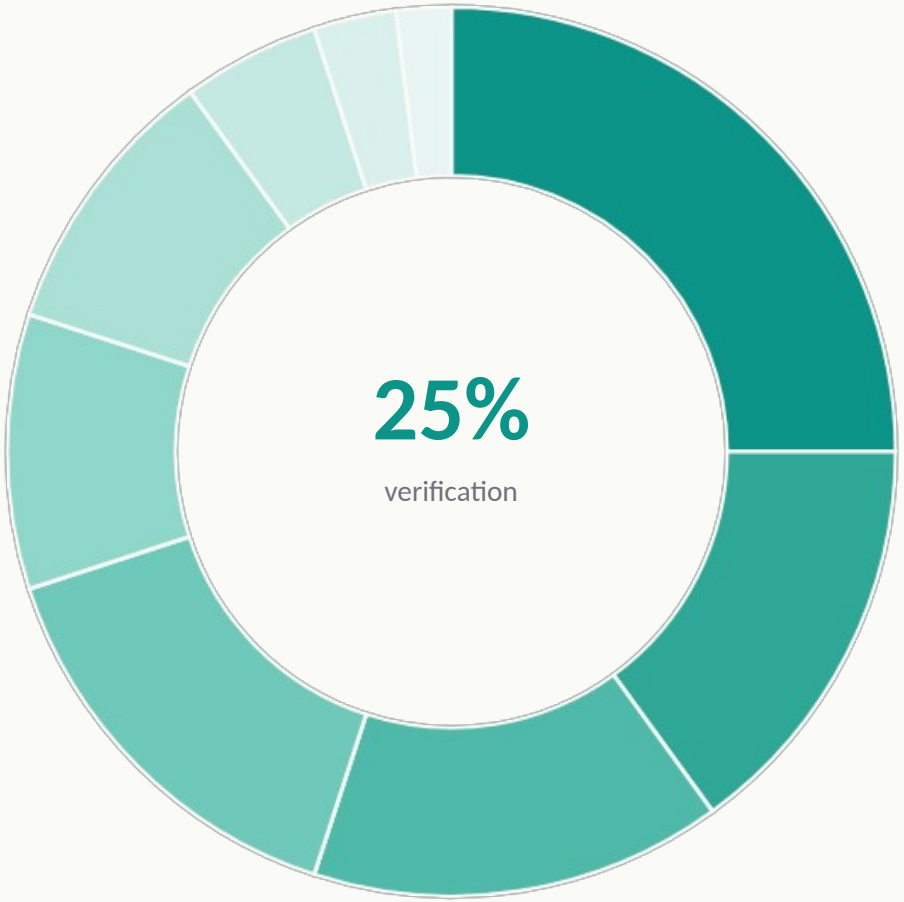
04

Try http:post

in strict enforcement mode, it raises before the side effect and the refusal is recorded

One variable changes between the runs, and it is the identity.

Trust must be continuous



A quarter of the score is whether the identity checks out.

The other eight watch behavior: uptime, action success, security alerts, compliance, isolation, age, drift, and the humans who supervise it. The same grant stops buying the same trust when the behavior changes.

Nine factors, weights in the open repo.

Untrusted contents is growing.

Start protecting your agent

Find what is running

```
npx hackmyagent detect
```

Wrap one agent

```
pip install aim-sdk
```

Break it on purpose

```
docker run -p 9000:9000 -p 7001-7021:7001-7021 opena2a/dvaa:0.9.2
```

Linkedin.com/in/abdelfane

opena2a.org/wearedevelopers