AIM.

Open Source Identity and Access Management for AI Agents.

Abdel Fane

CEO and Founder, OpenA2A · Executive Director, CSNP

OpenA2A

AGENDA

Five parts.

1	**The problem**	Why identity alone is not enough.
2	**What IAM teaches**	Forty years of identity engineering.
3	**The trust layer**	AIM and ARP. The architecture.
4	**Live demo**	Watch it defend in real time.
5	**What this unlocks**	Compliance, federation, contribution.

Q & A follows. 5 minutes.

WHO I AM

Abdel Fane.

CEO and Founder OpenA2A · Executive Director CSNP

20 years enterprise security

- Consulting at Booz Allen Hamilton and Protiviti
- In-house at Allstate, Grail
- Fortune 10 to small clinics

Agent security research contributions

- First public NemoClaw security assessment
- Findings disclosed via NVIDIA PSIRT, fixes shipped
- 8 PRs merged to OpenClaw (PR #9806)

01

PART 01

The problem.

Identity for agents is here. Trust is not.

Identity is the floor.

What sits above it?

OpenClaw: 310,000+ stars.

What happened	Why identity didn't help
'What Would Elon Do?' gamed the popularity algorithm. Thousands of installs. Cisco AI Defense found 9 vulns, 2 critical.	Identity tells you who published. It doesn't tell you what the skill does.
ClawHavoc. 800+ malicious skills seeded. ~20% of the registry. CVE-2026-25253: one-click RCE, CVSS 8.8.	Signing proves origin. It does not prove behavior.
Industry response. China restricted state agency use. NVIDIA shipped OpenShell at GTC 2026.	When identity is the only control, the only escalation is to abandon the platform.
We submitted 8 PRs. Including a 1,721-line skill safety scanner. PR #9806.	Patching is necessary. The architectural gap remains until trust becomes runtime.

What our fleet sees.

Window: April 12 – May 11, 2026 · Source: research.opena2a.org/reports

297,723

exposed AI services on the public internet

ARIAscout · May 12 sweep

Of those:

```
228,652   OpenClaw gateways     -13% MoM
206,571   honey-agent events    30 days
  9,037   unique fingerprints
     45%  returned attackers
155,674   targeted MCP          75%
```

The industry shipped identity. AGNTCY. Microsoft Agent Governance Toolkit. April to May 2026, OpenClaw gateways dropped 13%. The absolute number is still 228,652. Identity is available. The exposed surface persists.

THE LAYER ABOVE IDENTITY

Identity is point-in-time. Trust must be continuous.

The industry has built the floor:

- **AGNTCY**

Linux Foundation project. Cryptographically verifiable identity across organizational boundaries.

- **Microsoft Agent Governance Toolkit**

Released April 2026, MIT licensed. Sub-ms policy enforcement. All 10 OWASP agentic risks.

> *An agent verified 30 seconds ago could have visited a website with indirect prompt injection in those 30 seconds. The identity is still valid. The agent is now compromised. Continuous trust is the layer that catches this.*

AIM + ARP. The layer above identity.

02

PART 02

What IAM teaches.

Forty years of identity engineering. Three lessons.

EVERY ERA GETS ITS IDENTITY LAYER

Each transition took roughly a decade.

1961	First password (MIT CTSS)
1978	Public key cryptography
1988	Kerberos and SSO
1995	TLS and cryptographic server identity
2002	SAML and federated identity
2010	OAuth 2.0 and OpenID Connect
2014	WebAuthn and FIDO
2018	Zero Trust
2026	**AI agents — identity yes, trust no**

Not inventing a new field. Applying forty years of identity engineering to a new actor class.

LESSON 1

Cryptographic identity, not credentials.

Credentials (the old way)	Cryptographic identity
API keys	TLS certificates
Shared secrets	SSH keys
Bearer tokens	PKI
Stealable. Replayable.	The identity is the keypair.
Anyone holding the string is the user.	The private half never leaves the holder.

> Verification is signature math, not string match.

LESSON 2

Least privilege, not blanket access.

Blanket access	Capability-based
Admin password	Declared capabilities
Root token	Scoped per operation
Full IAM role	Each capability TTL-bounded

Salesloft-Drift & PocketOS:

- PocketOS AI deletes entire production DB in 9 seconds.
- Drift's OAuth token was stolen. Master key, broad scope. 700 organizations affected.
- Capability-scoped tokens would have stopped the same theft at the operational limit.

LESSON 3

Audit trails.

If it didn't happen in the audit log, it didn't happen.

Three requirements for an audit trail that holds up:

- Structured events. Not strings.
- Signed. Provenance verifiable.
- Append-only. No after-the-fact rewriting.

03

PART 03

The trust layer.

AIM and ARP. The architecture on top of identity.

AIM.

Agent Identity Management.

Continuous trust. Runtime enforcement. The layer above identity.

Apache 2.0

Self-host or AIM Cloud

Framework-agnostic (LangChain, CrewAI, AutoGen, raw MCP)

Three SDKs: Java, Python, TypeScript

AGENT IDENTITY

Keypair, not API key.

What AIM gives an agent instead of an API key.

```
agentId             aim_7f3a9c2e
publicKey           Ed25519
pqcPublicKey        ML-DSA-65
hybridModeEnabled   true
keyExpiresAt        2026-05-25T00:00:00Z
rotationCount       0
capabilities        ["flight:search", "flight:book"]
```

Hybrid keypair. Ed25519 today, ML-DSA-65 quantum-resistant alongside it. NIST IR 8547 sets the 2030 deprecation deadline. We are already past the migration date.

AGENT RUNTIME CONDUCT

5-step Fine-Grained Authorization (FGA).

Verify intent. Verify action. Verify behavior.

#	Step	What it catches
1	**Capability**	Does the agent hold the grant?
2	**Attribute**	Which fields, which actions?
3	**Context**	Drift score, scan verdict, alert count
4	**Chain**	Rate limits, sequence rules
5	**Intent**	NanoMind classification (HIGH risk only)

Steps 1-4 under 10 ms P99. Step 5 (NanoMind) up to 800 ms for HIGH risk, async for MEDIUM, skipped for LOW. Each step is scar tissue from a real attack class.

BEHAVIORAL TRUST SCORING

9 factors. Sums to 100.

#	Factor	Weight
1	**Verification status**	25%
2	**Uptime and availability**	15%
3	**Action success rate**	15%
4	**Security alerts (NanoMind-modulated)**	15%
5	**Compliance**	10%
6	**Execution isolation**	10%
7	**Age and history**	5%
8	**Drift detection**	3%
9	**User feedback**	2%

> Three factors require external attestation. An agent cannot self-attest into high trust.

MCP SERVER ATTESTATION

Every tool the agent connects to is itself an identity.

Primitive	What it does
Server identity	Every MCP server gets a cryptographic identity. AIM tracks which servers each agent is allowed to talk to.
Pre-connection scan	AIM scans the MCP server before allowing the agent to connect. Results land in scanSummary.
Drift detection	If an agent suddenly connects to a server it never registered, that is drift. Trust score drops.
Continuous re-attestation	7-day credential expiry forces re-scan. New CVEs in tool dependencies surface within the cycle.

HackMyAgent is the scanner that produces scanSummary. Apache 2.0. Documented separately.

SERVICENOW NOW ASSIST

Three vulnerabilities. 85% of the Fortune 500.

January 2026. Disclosed by AppOmni's Aaron Costello.

ServiceNow vulnerability	AIM primitive that closes it
Universal credential 'servicenowexternalagent' shared across all customers. No rotation. No per-tenant uniqueness.	Cryptographic identity
Email-only impersonation. Email + subdomain + universal credential. No password. No MFA.	Multi-factor agent auth
One prebuilt agent allowed 'create data anywhere in ServiceNow.' No scoping. No approval gates.	Capability-based access

HOW AIM CLOSES THE GAP

Each vulnerability maps to one primitive.

Cryptographic identity stops universal-credential reuse

Capability-based access stops 'create data anywhere'

Trust scoring detects the third unusual create-admin attempt before it succeeds

Audit trails make every attempt visible after the fact

Each is an ARC primitive. Each one would have stopped the exploit at a different layer.

04

PART 04

Live demo.

Flight booking agent. Prompt injection. AIM defending in real time.

Live demo.

Flight booking agent. Chat interface.

Live prompt injection.

AIM defending in real time.

WHAT YOU JUST SAW

Four things just happened.

What happened	How	Where
Identity verified	**Ed25519 signature checked locally**	Sub-5ms, no Registry call
Drift detected	**ASC score crossed threshold**	The moment injection landed
Action denied	**FGA Step 1 short-circuited**	Agent had no email:send
Audit trail captured	**Signed attestation written**	Transparency log, replayable

> That is ARC. Defense in depth, fail-fast at the cheapest layer.

THE INTELLIGENCE UNDERNEATH

Architecture is open. Intelligence is research.

What	What it does
Agent Threat Matrix	57 techniques across 9 tactics. Evidence-tiered: 16 observed, 38 lab-validated, 3 adapted. Submit to MITRE as D3FEND-style complement to ATT&CK and ATLAS.
Honeypot fleet	TrapMyAgent, AgentPwn, HoneyFinder, ARIAscout. 206,571 honey-agent events in 30 days. 9,037 unique fingerprints. 45% return rate.
NanoMind, fed by the fleet	Every attack the fleet observes becomes training data. Every defender running AIM + ARP gets smarter as the threat landscape evolves.

> *You can build the toolkit. You can't build the intelligence layer without the research infrastructure first.*

05

PART 05

What this unlocks.

Compliance. Open source. The work that's open.

COMPLIANCE MAPPING

Each framework maps to a primitive.

Framework	AIM primitive
SOC 2 CC6.3 — Privileged access	Capability tier classification
ISO 27001 A.9.2.3 — PAM	JIT capability grants with TTL
NIST SP 800-53 AC-6 — Least privilege	FGA attribute denials, row predicates
FedRAMP AC-2 / AU-9 — Audit integrity	Append-only signed audit log
EU AI Act Article 14 — Human oversight	Approval gates for high-risk operations
OWASP Agentic Top 10	Direct mapping to OASB 222 attack scenarios

Compliance is not a separate workstream. It falls out of the primitives.

OPEN SOURCE

Apache 2.0. Self-host. Fork. Audit.

100%

open source

Apache 2.0.

2

install paths

npm or Docker

3

SDKs

Java, Python, TypeScript

AIM Cloud at aim.opena2a.org if you do not want to run infrastructure.

WHAT'S COMING NEXT

ATX. Agent Trust eXtension.

The credential layer the architecture needs. Currently in spec development.

- Published at IEEE SVCC 2026 (San Jose, June 10 to 12)
- Reference implementation tracks Agent Trust Protocol (ATP) v1.0.0-rc1
- Public record: threats.opena2a.org
- RFC 6962-compatible transparency log · federated CRL · 5-minute revocation

> *DIDs are point-in-time. AIM and ARP make trust continuous.*
> *ATX is the credential format that assumes continuous re-verification.*
> *The two are designed together.*

Try AIM.

aim.opena2a.org

```
npx opena2a-cli review
```

github.com/opena2a-org · research.opena2a.org · Apache 2.0.

Q & A

Receipts: research.opena2a.org

opena2a.org · github.com/opena2a-org · abdel@opena2a.org

Thank you.

Connect on LinkedIN

OpenA2A.org