OpenA2A

Observing the Observers.

Bringing OpenTelemetry to Autonomous AI Agents.

Abdel Fane

CEO and Founder, OpenA2A · Observability Summit NA 2026 · Ballroom B

WHAT WE ARE SEEING IN THE WILD

Agents are already under attack.

OpenA2A runs a honeypot fleet of AI agents. 30 days of behavioral telemetry, April to May 2026:

206,571

attacker events

captured against honey agents in 30 days

75%

targeted MCP

three of every four events hit the agent protocol layer

45%

attackers returned

came back across multiple sessions. Reconnaissance, not noise.

Standard telemetry does not see the authorization layer.

Behavioral Threat Report Issue 1. ARIA autonomous research system, ed. Abdel Fane. research.opena2a.org

AGENDA

Four parts.

1	**The problem**	Agents are under attack. The telemetry is blind to it.
2	**The proposal**	Nine attributes. Issue #180. Producer-emitted, not normative.
3	**Detection**	What standardized names unlock against returning attackers.
4	**Demo and asks**	Two agents, same nine attributes. Then what to do.

Q & A follows. Four minute live demo at the end.

WHO I AM

Abdel Fane.

CEO and Founder, OpenA2A · Contributor, OTel SemConv SIG GenAI

20 years enterprise security

- Booz Allen Hamilton, Protiviti, Allstate, Grail
- Cloud security and DevSecOps
- Fortune 10 to small clinics

Agent observability work

- Issue #180: agentic authorization SemConv
- AIM backend emits the FGA authorization spans
- Filed last week. Public reference impl on GitHub.

THE ECOSYSTEM IS RESPONDING

The landscape filled in.

What shipped between November 2025 and today:

Project	What it covers
AGNTCY (joined LF, July 2025)	Identity, discovery, messaging, observability for cross-vendor agents.
Agentgateway (joined LF, August 2025)	Gateway RBAC and visibility for MCP servers.
AAIF (formed December 2025)	LF umbrella. AWS, Anthropic, Google, Microsoft, OpenAI, and others.
OTel GenAI SemConv v1.37+	LLM call attributes: model, tokens, provider. In the official spec.
OTel agentic SemConv (Issue #35)	Tasks, actions, agents, tools, memory. Active draft work.
MCP SemConv (PRs #2043, #2083)	MCP server identity attributes. Shipped.

THE GAP

What hasn't been standardized yet.

The attacks land in three dimensions nobody has standardized:

1 Authorization decision attributes

What capability was invoked. By which agent. With what outcome. The decision path is invisible to GenAI and agentic conventions today.

2 Behavioral signals as decision inputs

Trust scores, drift signals, scan verdicts. Producer-emitted values that feed authorization decisions. Not normative. Not subjective. Just attributes the producer chose to publish.

3 FGA decision path

Multi-step authorization spans. Which step denied, why, what the outcome was. Fine-grained authorization observability.

WHY THIS PROPOSAL IS DIFFERENT

Observability sees what happened. Not who.

What a trace shows today

A capability was invoked. A tool was called. An anomaly occurred at 14:32.

The event is recorded. But "agent" in current telemetry is a label, not a verified identity. Anyone can write it.

What it cannot answer

Which verified agent. Under what capability grant. Issued by whom. Expiring when.

And: is this the same agent that was here last session? You cannot say an attacker returned without verified identity.

The nine attributes in this proposal are anchored to a cryptographic agent identity. That anchor is what makes them different from a label in a span.

Observability records the action. Identity proves the actor.

01

PART 01

The proposal.

Nine attributes. Producer-emitted, not normative.

ISSUE #180

Nine attributes. Three groups.

Identity

```
agent.id
agent.public_key.algorithm
```

Who is acting. How they sign.

Action and decision inputs

```
agent.capability
agent.trust_score
agent.drift_score
agent.scan_verdict
```

What they are doing. What signals the producer published.

FGA decision path

```
fga.step
fga.outcome
fga.denied_by
```

Which authorization step. What it decided.

All nine attribute names are emitted by the AIM backend FGA path. Reference implementation linked from Issue #180.

THE FRAMING THAT MATTERS

Producer-emitted. Not normative.

The trap

An earlier proposal at semantic-conventions #3582 was closed because reviewers called trust scores subjective. They said scores like that do not belong in a SemConv registry.

That critique is correct if you frame trust_score as a normative value everyone must compute the same way.

The framing

trust_score and drift_score are producer-emitted decision inputs. Like a sampling decision.

The producer computes the score using whatever method makes sense for their domain. The spec only standardizes the attribute name, type, and range so downstream observers can correlate.

Same role as gen_ai.usage.input_tokens. The producer publishes the value. Consumers correlate it. The spec does not mandate how it was computed.

WHAT IT LOOKS LIKE ON THE WIRE

One parent span. The decision path as children.

Real span tree from a trace verified in Tempo today:

```
fga.authorize                                [parent]
├── agent.id                    "38176c67-..."
├── agent.public_key.algorithm  "ed25519"
├── agent.capability            "file:read"
├── agent.trust_score           0.92
├── agent.drift_score           0.03
├── agent.scan_verdict          "CLEAN"
├── fga.outcome                 "DENY"
├── fga.denied_by               "capability_check"
│
└── fga.capability_check                     [child]
    ├── agent.id                "38176c67-..."
    ├── agent.capability        "file:read"
    ├── fga.step                "capability_check"
    └── fga.allowed             false
```

HOW THIS FITS

Three layers. None redundant.

Layer	Owner	What it answers
Identity	AGNTCY	Cryptographic identity. Who is acting across organizational boundaries.
LLM call	OTel GenAI SemConv v1.37	What the model did. gen_ai.request.model, usage.input_tokens, etc.
Authorization	Issue #180 (this proposal)	What the runtime decided. agent.capability, fga.outcome, decision inputs.

> *Identity tells you who. The model layer tells you what the LLM produced. The authorization layer tells you what the runtime decided.*

02

PART 02

Detection.

Standardized names enable cross-vendor queries.

ANOMALY DETECTION

Two patterns. Neither mandated.

Approach	When it works, when it does not
Statistical baselines	Histograms over agent.trust_score and agent.drift_score by capability. Alert when an agent moves more than N standard deviations from its baseline. Cheap. Explainable. Fails when behavior changes legitimately.
ML-driven classification	Train a classifier on the standardized attributes plus span features. Detects patterns humans would not write rules for. Better recall on novel attacks. Requires labeled data and retraining. Standardized names make portable models possible.

> *The spec stays agnostic. The producer publishes the decision inputs. Consumers correlate however makes sense for their environment.*

WHAT YOU CAN ASK

Queries that names unlock.

1. Which capabilities are denied most often, and by which step?
2. Do agents with lower trust_score produce more downstream errors?
3. Which scan_verdict values predict the highest deny rates?
4. How does drift_score correlate with capability success over time?
5. Which agents concentrate on capabilities outside their declared scope?

> *None of these queries work today without writing custom translations for each vendor's proprietary attribute names.*

03

PART 03

Demo.

AIM backend. LangChain bridge. Same Tempo dashboard.

Same spec. Two agents.

Segment 1 AIM backend emits the nine attributes on a real fga.authorize span.

Segment 2 Minimal LangChain agent emits the same nine attributes via the bridge.

Watch it for the attribute names.

WHY THIS MATTERS NOW

History rhymes.

Before OTel

Datadog. New Relic. AppDynamics. Dynatrace. Others.

Every APM vendor shipped its own SDK. Different attribute names for the same concepts. Migration cost was a tax on every engineering team.

The industry built OpenTelemetry to fix this. It took years.

Now, for agents

Datadog, New Relic, Dynatrace, and others are each shipping proprietary agent telemetry.

None of it interoperates.

We are about to repeat the same pattern. The OTel project exists to prevent it.

Standardize now while the spec is still being defined.

WHAT TO DO WITH THIS

Three asks.

1 Review the proposal

Comment on Issue #180 at open-telemetry/semantic-conventions-genai. Attribute names especially need your eyes.

2 Try the bridge

github.com/opena2a-org/otel-semconv-agent-identity. Five lines of instrumentation on your LangChain agent. Send us a trace.

3 Help co-author the upstream PR

If you are active in OTel SemConv governance, find me after this talk. We need a co-author.

Standardize before the lock-in.

Abdel Fane · abdel@opena2a.org · opena2a.org