Agent Operation Authorization

https://datatracker.ietf.org/doc/html/draft-liu-agent-operation-authorization-01

Dapeng Liu / Suresh Krishnan

Problem Statement & Design Goals

Current OAuth Limitations in AI Agent Context:

1. Permission ≠ Behavior (The Injection Risk)

- **Current OAuth**: Grants broad Permissions (e.g., scope=write).
- **The Risk**: Since LLMs struggle to distinguish **Data vs. Instruction**, a **Prompt Injection** can trick the Agent into executing malicious actions that are still technically within the granted Permission.
- **Gap**: A mechanism to **bind specific behavioral policies** with OAuth authorization.

2. The Agent Delegation Challenge (Intent Decoupling)

- **Hallucination in Parsing intent:** The Agent may misinterpret the user's natural language intent.
- **Missing Contextual Binding:** Current OAuth proves who the user is, but not **which Agent is acting which user and what specific task** was authorized.
- **Gap:** Lack a cryptographic link for "Agent X acting for User Y to do Task Z".

Design Goals: Bind behavioral Policy +Agent Delegation + User Intent Evidence into the Access Token.

Solution: Agent Operation Authorization (AOA)

AOA Solution: Policy-Based Authorization with AS-signed evidence binding **User + Agent + Operation**

The agent_operation_proposal Mechanism

Core Concept:

- *A structured, machine-readable proposal generated by the Agent, serving as the **binding contract** between User, Agent, and Operation.*

1. Explicit Tripartite Binding (The "Who-What-Which" Gap Solved)

AOA explicitly cryptographically binds three elements:

- ✓ **Which User**: Identifies the Resource Owner (user identity).
- ✓ **Which Agent**: Identifies the specific Agent instance (agent identity).
- ✓ **What Specific Behavior**: Defines the exact operation policy constraints (e.g., amount <= 50, time == ”Nov. 11 night").

Agent → [Structured Proposal] → AS validates User + Agent + Intent + Policy → Signed Token → RS → [Enforces Policy]

Security Model & Interoperability

Trust Boundaries:

- LLM: Untrusted (Prompt Injection/Hallucination possible).
- Agent Runtime: Trusted runtime responsible for signing requests.
- AS: The sole Witness capable of issuing evidence and extending chains.

Ecosystem Fit:

- OAuth: Uses PAR, JWT, standard error codes.
- WIMSE: Leverages Workload Identity for Agent authentication.
- OPA: Standard engine for Rego policy enforcement.

Open Source Implementation:

- **Open Agent Auth**
- IETF 125 Hackathon

Open Agent Auth

https://github.com/alibaba/open-agent-auth

Summary

Value Proposition:

- User Identify + Agent Identify + behavioral policies based Authorization
- Solves AI Hallucination/Over-privilege risks.
- Provides semantic audit trails (Human -> Agent -> RS).
- Unified model for Single & Multi-Agent orchestration.

Questions for the WG:

- **Seeking Consensus on the Problem**
  - ✓ *Does the WG agree that Agent-specific risks (Prompt Injection, Hallucination) create a critical gap in current OAuth (Scope-based) for Agent scenarios?*
- **Validating the Direction**
  - ✓ *Is shifting from Scope-Based to Policy-Based Authorization (with AS-signed evidence) the right architectural direction to solve this?*
- **Proposed Next Steps**
  - ✓ *Form a small Design Team to refine the solution.*

Thanks!