

AI Agent Identity, Access and Authorization Architecture

A Zero-Trust Enterprise Architecture for Agent Identity, Authentication, Delegation, Least Privilege, Tool Authorization, and Governed Autonomous Action

Sanjeeve Kumar Gajadi*

*(HCL Technologies, Bengaluru, Karnataka, India)

Email: sanjeevekumar.g@hcltech.com

Abstract:

AI agents increasingly act as digital workers that invoke APIs, access enterprise data, delegate tasks, and execute transactions across business systems. Treating agents as anonymous application processes creates unacceptable security and accountability risk because organizations cannot reliably determine who or what acted, under whose authority, with which privileges, or for what purpose. This paper proposes an AI Agent Identity, Access and Authorization Architecture based on Zero Trust, workload identity, least privilege, delegated authority, policy-based access control, short-lived credentials, transaction-level authorization, and continuous verification. The architecture distinguishes human, application, agent, sub-agent, and tool identities; defines trust boundaries and delegation chains; and establishes controls for secrets, impersonation, token exchange, tool permissions, data entitlements, approval gates, and revocation. It also addresses multi-agent systems, high-risk transactions, audit evidence, emergency shutdown, and lifecycle governance. The proposed architecture enables enterprises to deploy autonomous agents with strong security, traceability, and business accountability.

Keywords — AI Agent Identity, Agentic AI Security, Zero Trust, Workload Identity, Authorization, Least Privilege, Delegated Authority, Multi-Agent Systems, Identity Governance, Enterprise Architecture

I. INTRODUCTION

AI agents can perform actions that were previously executed by employees or application services. Identity architecture must therefore distinguish agent actions from user actions and maintain clear accountability.

An agent should never inherit broad authority merely because it operates on behalf of a privileged user or application.

II. IDENTITY MODEL

The architecture defines separate identities for humans, applications, agents, sub-agents, orchestration services, tools, and external providers.

Each identity should have a unique identifier, owner, purpose, lifecycle state, trust level, permissions, and auditable relationship to the business process it supports.

III. AUTHENTICATION ARCHITECTURE

Agents should use strong workload authentication based on certificates, managed identities, signed tokens, or equivalent mechanisms rather than embedded passwords.

Credentials should be short-lived, rotated automatically, scoped to environment, and protected by centralized secrets or key management.

IV. AUTHORIZATION MODEL

Authorization should combine role-based, attribute-based, policy-based, and transaction-level controls according to business risk.

Tool access, data access, financial limits, customer-impacting actions, and administrative operations should be explicitly authorized rather than inferred from model output.

V. DELEGATION AND IMPERSONATION

Delegation must preserve the chain of authority from human or business process to orchestrator, agent, sub-agent, and tool.

Delegated tokens should be constrained by purpose, scope, time, transaction amount, data classification, and permitted actions.

VI. LEAST PRIVILEGE AND JUST-IN-TIME ACCESS

Agents should receive only the permissions necessary for a specific task and duration. Persistent broad privileges increase the impact of prompt injection, compromised models, or logic errors.

Just-in-time access and dynamic policy evaluation can reduce standing privilege while preserving operational flexibility.

VII. TOOL AND API AUTHORIZATION

Every tool exposed to an agent should have an explicit contract, identity requirement, authorization policy, input validation, rate limit, and audit trail.

High-risk tools should support additional approval steps, dual control, transaction limits, or restricted operating windows.

VIII. DATA ACCESS CONTROLS

Agent data access should preserve source-system entitlements and classification rules. Retrieval services should not expose information simply because the agent can technically reach the source.

Sensitive data may require masking, purpose limitation, jurisdiction controls, and prevention of onward disclosure to external models.

IX. MULTI-AGENT IDENTITY

Multi-agent systems require clear identity propagation across delegation chains. Sub-agents should not silently inherit all privileges of parent agents.

Each delegation event should record issuer, recipient, scope, expiration, task purpose, and policy conditions.

X. CONTINUOUS VERIFICATION AND REVOCATION

Zero Trust requires continuous evaluation of identity, context, behavior, device or workload posture, and transaction risk.

Organizations need immediate revocation capabilities for agents, credentials, tools, sessions, and delegated permissions when abnormal behavior or incidents occur.

XI. AUDITABILITY AND NON-REPUDIATION

Audit records should link each agent action to identity, authorization decision, delegated authority, tool invocation, data accessed, and business outcome.

Strong traceability supports incident investigation, regulatory evidence, segregation-of-duties review, and accountability.

XII. LIFECYCLE GOVERNANCE

Agent identities should be created through approved workflows, periodically reviewed, recertified, and retired when use cases end or ownership changes.

Orphaned agent identities and unused credentials should be treated as high-priority security risks.

XIII. CONCLUSION

AI agents require identity and authorization controls comparable to privileged digital workers rather than generic application processes.

The proposed architecture provides a Zero-Trust foundation for secure autonomous action through explicit identity, scoped authority, least privilege, delegation controls, and continuous verification.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, AI governance communities, security professionals, standards bodies, and technology researchers whose published work has informed the concepts synthesized in this paper.

REFERENCES

- [1] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.

- [2] National Institute of Standards and Technology, AI RMF Generative Artificial Intelligence Profile, 2024.
- [3] ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system.
- [4] ISO 31000:2018, Risk management — Guidelines.
- [5] ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems.
- [6] National Institute of Standards and Technology, SP 800-207: Zero Trust Architecture, 2020.
- [7] The Open Group, TOGAF Standard, 10th Edition, 2022.
- [8] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
- [9] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
- [10] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
- [11] OECD, Recommendation of the Council on Artificial Intelligence.
- [12] IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.