

AI Agent Accountability and Digital Evidence Management Framework

A Comprehensive Enterprise Framework for Establishing Agent Ownership, Decision Responsibility, Evidence Integrity, Auditability, and Defensible Digital Records

Sanjeeve Kumar Gajadi*

*(HCL Technologies, Bengaluru, Karnataka, India)

Email: sanjeevekumar.g@hcltech.com

Abstract:

AI agents can initiate recommendations, decisions, communications, transactions, and multi-step workflows across enterprise systems. As autonomy increases, organizations must establish who is accountable for agent configuration, delegated authority, business outcomes, and failures. They must also retain reliable digital evidence demonstrating what the agent knew, what it attempted, which controls were applied, and who approved critical actions. This paper proposes an AI Agent Accountability and Digital Evidence Management Framework that combines ownership models, decision rights, agent identity, delegation records, audit trails, evidence integrity, retention, legal hold, access control, chain of custody, and investigation workflows. The framework defines accountability across business owners, agent owners, model owners, data owners, platform teams, and human approvers. It also addresses evidence from prompts, retrieved context, tool calls, policy decisions, approvals, outputs, exceptions, and downstream actions. The proposed approach enables enterprises to operate AI agents with stronger governance, audit readiness, incident response, and defensible accountability.

Keywords — AI Agent Accountability, Digital Evidence, Agentic AI, Auditability, AI Governance, Chain of Custody, Decision Rights, AI Forensics, Digital Records, Enterprise Architecture

I. INTRODUCTION

Autonomous execution can blur responsibility if organizations treat AI agents as neutral technical components.

Accountability requires explicit ownership of the agent, business purpose, delegated authority, data use, controls, and resulting outcomes.

II. ACCOUNTABILITY MODEL

Key roles include business owner, agent owner, model owner, data owner, platform owner, security owner, human approver, and audit or compliance functions.

Roles should distinguish who designs, operates, approves, monitors, and accepts risk.

III. DECISION RIGHTS

Decision rights should define which actions agents may execute autonomously, which require human approval, and which are prohibited.

Thresholds can depend on financial value, customer impact, safety, regulation, data sensitivity, and reversibility.

IV. AGENT IDENTITY AND DELEGATION

Each agent should have a unique identity and a traceable relationship to the user, process, or business authority that initiated the task.

Delegation records should preserve scope, duration, purpose, and downstream sub-agent authority.

V. DIGITAL EVIDENCE MODEL

Evidence should include prompts, context, retrieved sources, model version, policies, tool calls, authorization decisions, approvals, outputs, timestamps, and transaction results.

Evidence must be sufficient to reconstruct material actions without collecting unnecessary sensitive data.

VI. EVIDENCE INTEGRITY

Critical records should use immutable or append-only storage, integrity hashes, synchronized timestamps, and controlled write permissions.

Integrity controls support non-repudiation and increase confidence during investigation or audit.

VII. CHAIN OF CUSTODY

Evidence handling should record who collected, accessed, exported, transformed, or disclosed records during investigations.

Chain-of-custody controls are especially important when evidence may support legal, regulatory, disciplinary, or contractual processes.

VIII. RETENTION AND LEGAL HOLD

Retention should reflect regulatory, contractual, operational, privacy, and investigation requirements.

Legal-hold mechanisms should prevent deletion of relevant agent evidence when disputes or investigations arise.

IX. PRIVACY AND ACCESS CONTROL

Evidence repositories may contain confidential prompts, customer data, credentials, or business information and require strict access controls.

Role-based access, masking, encryption, and purpose limitation can reduce secondary privacy risk.

X. INVESTIGATION AND AUDIT

Investigators should be able to reconstruct a sequence from initial instruction through delegation, policy checks, human approval, tool execution, and final outcome.

Standard evidence packages can accelerate incident response, model-risk review, and external audit.

XI. ACCOUNTABILITY METRICS

Metrics include percentage of agents with named owners, evidence completeness, unresolved ownership gaps, missing approvals, expired delegations, and investigation response time.

Repeated accountability failures should trigger governance remediation and architecture review.

XII. CONCLUSION

AI Agent Accountability and Digital Evidence Management Framework establishes a defensible governance foundation for autonomous enterprise action.

By combining clear ownership, bounded authority, traceable decisions, and integrity-protected evidence, organizations can scale agentic AI while preserving accountability and audit readiness.

ACKNOWLEDGMENT

The author acknowledges the contributions of enterprise architecture practitioners, AI governance specialists, standards bodies, security professionals, data leaders, and technology researchers whose published work 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] The Open Group, TOGAF Standard, 10th Edition, 2022.
- [7] ISACA, COBIT 2019 Framework: Governance and Management Objectives, 2019.
- [8] OWASP Foundation, OWASP Top 10 for Large Language Model Applications.
- [9] MITRE, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.
- [10] OECD, Recommendation of the Council on Artificial Intelligence.
- [11] SAP SE, SAP Business Technology Platform Documentation.
- [12] SAP SE, SAP Business Data Cloud and SAP Datasphere Documentation.