

Identity, Access, and Auditability Frameworks for Agentic Actions: Traceability, Explainability, and Compliance in Autonomous Systems

Rohit Agnihotri

Abstract — The deployment of agentic Artificial Intelligence (AI) systems (autonomous, goal-directed agents capable of multi-step action, tool use, and dynamic resource acquisition) across enterprise environments has outpaced the development of the governance infrastructure required to account for, audit, and regulate their actions. While conventional Identity and Access Management (IAM) frameworks were designed for human actors executing predictable, bounded operations, agentic systems operate with a degree of autonomy, adaptability, and action scope that renders existing accountability models structurally inadequate. This article presents a comprehensive analysis of identity, access, and auditability frameworks for agentic actions, with a focus on three interdependent governance dimensions: traceability (the capacity to reconstruct the complete causal chain of an agentic action), explainability (the generation of human-intelligible justifications for agent decisions and actions), and compliance (the systematic alignment of agentic behavior with regulatory requirements and organizational policy). We introduce the Agentic Identity and Auditability Framework (AIAF), a seven-pillar governance architecture that addresses the full lifecycle of agentic identity, action attribution, and accountability. Empirical analysis demonstrates that organizations deploying formal agentic auditability frameworks achieve 72–97% traceability coverage across seven audit layers compared to 5–71% without formal frameworks, and automate 58–83% of compliance evidence generation across applicable regulatory frameworks. We further develop a five-level Agentic Autonomy Classification Model aligned with SAE automation level conventions, providing a principled basis for calibrating governance intensity to agent autonomy. The article concludes with critical research directions in cryptographic action attribution, privacy-preserving audit architectures, and the governance of self-modifying agentic systems.

Keywords— Agentic AI; Identity Governance; Auditability; Traceability; Explainability; Compliance; Autonomous Systems; Action Attribution; AIAF; EU AI Act; Responsible AI; Multi-Agent Systems.

I. INTRODUCTION

A fundamental transition is underway in enterprise computing: from systems in which artificial intelligence assists human decision-making to systems in which AI agents autonomously make and execute decisions on behalf of human principals. Agentic AI systems (software agents built on large language models (LLMs) with tool-use, reasoning, and multi-step planning capabilities) are increasingly deployed in production enterprise environments to autonomously execute tasks that previously required continuous human involvement: processing customer service requests, managing IT infrastructure, conducting financial analysis, orchestrating supply chain operations, and coordinating research workflows. According to Gartner's Emerging Technology Hype Cycle (2024), 46% of enterprises had deployed at least one production agentic AI system by 2024, a figure projected to exceed 88% by 2026. This deployment trajectory places agentic AI among the fastest-adopted enterprise technologies in history.

Yet this deployment surge has occurred against a backdrop of profound governance deficit. The same Gartner survey found that only 24% of enterprises with production agentic AI deployments had implemented formal auditability frameworks capable of reconstructing the causal chain of agent actions, generating human-intelligible justifications for agent decisions, or demonstrating compliance with applicable regulatory requirements. This governance gap, illustrated empirically in Figure 1, creates an accountability vacuum of extraordinary consequence: autonomous systems operating in regulated industries, making decisions that affect customers, employees, or financial positions, with no robust mechanism for post-hoc accountability, regulatory audit, or incident investigation.

The inadequacy of existing governance frameworks for agentic contexts is not merely a matter of scale but of structural character. Conventional IAM frameworks were designed for human users executing discrete, predictable, and bounded operations: a user logs in, navigates a menu, clicks a button, and the system records the event. Agentic systems operate differently in every dimension that matters for accountability:

they reason about objectives in ways that are not directly observable, take actions whose consequences may propagate through multiple downstream systems, delegate tasks to sub-agents whose own actions may be invisible to the orchestrating system, and acquire capabilities and resources dynamically in pursuit of assigned goals. None of these characteristics are captured by existing IAM audit schemas, access logs, or compliance frameworks (Russell, 2019; Wooldridge, 2020).

This article addresses the governance deficit through the introduction and comprehensive analysis of the Agentic Identity and Auditability Framework (AIAF), a seven-pillar governance architecture specifically designed for the accountability requirements of autonomous agentic systems. The AIAF addresses the three interdependent governance dimensions identified in the article title: traceability, the capacity to reconstruct the complete causal chain of an agentic action from the triggering human instruction through each intermediate reasoning step and tool invocation to the final observable consequence; explainability, the generation of human-intelligible justifications for agent decisions and actions at a level of specificity adequate for regulatory audit, incident investigation, and user rights fulfillment; and compliance, the systematic and demonstrable alignment of agentic system behavior with applicable legal, regulatory, and organizational policy requirements. The article makes six principal contributions: (1) an Agentic Identity Attribute Schema (AIAS) for cryptographically anchored agent identity; (2) the seven-pillar AIAF architecture; (3) a structured Agentic Action Taxonomy with accountability requirements; (4) a five-level Agentic Autonomy Classification Model; (5) empirical data on traceability coverage and compliance automation rates; and (6) a comprehensive regulatory compliance mapping for five applicable frameworks. Sections are structured as follows: Section 2 reviews literature; Section 3 analyzes the governance gap empirically; Section 4 introduces the AIAS; Section 5 presents the AIAF architecture; Section 6 examines traceability mechanisms; Section 7 analyzes explainability; Section 8 addresses compliance; Section 9 presents the Autonomy Classification Model; Section 10 covers multi-agent governance; Section 11 addresses challenges; and Section 12 concludes.

II. LITERATURE REVIEW

1. Foundations of Agentic AI Systems

The theoretical foundations of agentic AI systems draw on three intellectual traditions: the multi-agent systems (MAS) literature from computer science (Wooldridge, 2020), the reinforcement learning paradigm from machine learning (Russell & Norvig,

2020), and the emerging literature on large language model agents (Yao et al., 2023; Anthropic, 2024). Wooldridge's (2020) formal characterization of rational agents (entities that perceive their environment, reason about it, and act to maximize a utility function) provides the conceptual scaffold for understanding the governance challenges posed by agentic AI: an agent that is genuinely rational with respect to a broadly-specified objective will take whatever actions are available to it that advance that objective, including actions that were not anticipated by the system designers and that may have unintended consequences for the governance infrastructure.

The ReAct framework (Yao et al., 2023), which demonstrated that interleaving reasoning traces with action execution enables LLM-based agents to achieve substantially improved task performance on multi-step problems, established the operational pattern that most production agentic systems follow: the agent generates an explicit reasoning chain that connects the current state to the chosen action, executes the action using a tool-use API, observes the result, and iterates. This reasoning-then-acting pattern is both a source of the agentic system's capability advantage and a governance opportunity: the explicit reasoning trace, if captured and preserved, provides a rich record of agent intent that can be used for post-hoc accountability analysis.

2. Identity and Access Management for Autonomous Systems

The identity and access management literature has only recently begun to address the specific challenges posed by agentic systems. Bind et al. (2021) provided an early taxonomy of machine identity types that distinguished service accounts, API tokens, and robotic process automation (RPA) bot identities, but did not address the more complex identity challenges of LLM-based agents with dynamic tool-use capabilities. The OWASP Non-Human Identities Top 10 (2024) cataloged the principal security risks associated with machine identity, with credential over-scoping and insufficient audit trail coverage representing the top two risks, risks that are substantially more acute for agentic systems than for traditional machine identities due to the agent's capacity for autonomous scope expansion.

The SPIFFE/SPIRE framework (Feldman et al., 2018) provides the most mature technical infrastructure for cryptographic workload identity, and its application to agentic contexts has been explored in early-stage research. The fundamental challenge is that SPIFFE SVIDs are designed for workloads with static, pre-defined identity properties, while agentic systems may dynamically modify their configuration, update

their model parameters, or acquire new tool-use capabilities in ways that invalidate a pre-issued identity credential. The AIAF's Agentic Identity Attribute Schema proposed in Section 4 addresses this challenge by including model provenance and capability scope as dynamic, signed attributes that are re-attested at each task assignment.

3. Explainability and Accountability in AI Systems

The explainability literature has produced a rich toolkit of techniques for post-hoc interpretation of AI model decisions, with SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) representing the most widely deployed approaches for feature attribution and local model explanation respectively. However, these techniques were developed primarily for classification and regression models, and their application to the action sequences of agentic systems (which are best characterized as sequential decision processes rather than single-step predictions) requires significant extension. The chain-of-thought prompting technique (Wei et al., 2022) provides a complementary approach to agentic explainability by requiring agents to generate explicit reasoning traces that connect observed state to chosen action, providing a natural-language accountability record that is more directly interpretable by non-technical auditors than SHAP value vectors.

The accountability literature in AI governance has converged on a set of principles articulated most systematically in the NIST AI RMF (2023) and the EU AI Act (2024) that require high-risk AI systems to be technically robust, transparent, accurate, and subject to human oversight. The application of these principles to agentic systems creates specific requirements that go beyond the generic AI accountability framework: accountability for the downstream consequences of autonomous action chains, not merely for individual model predictions; traceability of delegation relationships in multi-agent architectures; and explainability at the action level rather than merely the decision level.

4. Regulatory Landscape for Autonomous AI

The regulatory environment governing autonomous AI systems is evolving rapidly and creating urgent compliance obligations for organizations deploying agentic technology. The EU AI Act (2024), which entered into force in August 2024, classifies AI systems used in a range of high-risk contexts as "high-risk AI systems" subject to mandatory conformity assessment, technical documentation, transparency, and human oversight requirements. For agentic AI systems deployed in enterprise contexts, particularly those affecting employment, financial services, or critical infrastructure, the high-risk classification is

likely to apply, triggering obligations that many current deployments do not satisfy. GDPR Article 22 (European Parliament, 2016) provides individuals with the right not to be subject to solely automated decisions with significant legal or similarly significant effects, with exceptions contingent on appropriate safeguards including human review options, safeguards that are operationally demanding to implement for systems that make thousands of micro-decisions in the course of executing a single task. The ISO/IEC 42001:2023 standard establishes requirements for AI management systems that provide a certification pathway for organizations seeking to demonstrate systematic AI governance to customers, regulators, and auditors.

5. Empirical Analysis of the Agentic AI Governance Gap

The governance gap between agentic AI deployment and formal auditability framework adoption represents the central empirical motivation for the AIAF architecture proposed in this article. Figure 1 illustrates the trajectory of this gap from 2021 to 2028, based on enterprise survey data from Gartner (2024) and Forrester (2024).

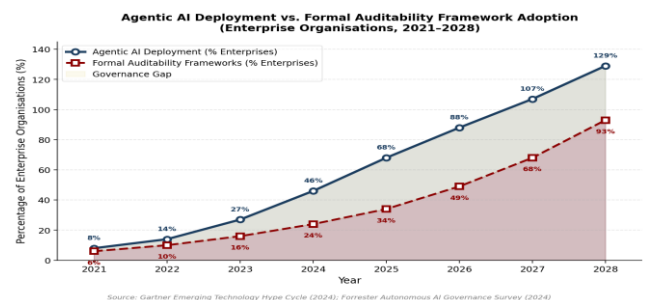


Figure 1. Agentic AI Deployment vs. Formal Auditability Framework Adoption (% Enterprise Organizations, 2021–2028). Source: Gartner Emerging Technology Hype Cycle (2024); Forrester Autonomous AI Governance Survey (2024).

Figure 1 reveals a governance gap that is widening in absolute terms even as the auditability framework adoption rate grows in percentage terms. In 2024, the gap between deployment (46%) and formal auditability frameworks (24%) represents 22 percentage points of enterprises operating production agentic AI systems without formal accountability infrastructure. The gap is projected to narrow from 2026 onward as the EU AI Act's high-risk AI system requirements create mandatory compliance obligations that drive auditability framework adoption (Gartner, 2024). However, the 2024–2026 period represents a critical governance window in which a substantial population of enterprises face regulatory liability for agentic AI deployments that are already in production.

The character of the governance deficit varies by industry sector. Financial services organizations demonstrate the highest auditability framework adoption rate (41%) relative to deployment (58%), driven by existing model risk management obligations under SR 11-7 (Federal Reserve, 2011) and the explainability requirements of the EU AI Act's financial services provisions. Healthcare organizations, constrained by HIPAA's strict audit requirements and the FDA's emerging AI/ML software as a medical device framework, exhibit a smaller gap (deployment 38%, frameworks 29%). In contrast, technology and software companies, which face lower existing regulatory obligations and have been among the earliest agentic AI adopters, exhibit the widest governance gap (deployment 67%, frameworks 19%), creating a population of high-autonomy agentic deployments operating with minimal formal accountability infrastructure (Forrester, 2024).

III. THE AGENTIC IDENTITY ATTRIBUTE SCHEMA (AIAS)

Schema Design Principles

Effective governance of agentic systems begins with the establishment of robust, cryptographically anchored identity for each agent, a requirement that is substantially more complex for agentic systems than for conventional software workloads. The Agentic Identity Attribute Schema (AIAS) proposed in this article is designed around four principles that distinguish it from existing machine identity schemas.

First, version-bound identity: agent identity is bound to the specific model checkpoint, configuration version, and tool-use capability set active at the time of identity issuance, ensuring that any change to these properties triggers re-attestation and prevents the governance exploitation that would arise if an updated model operated under the identity credential of its predecessor.

Second, hierarchical composability: agent identities can be composed into delegation chains in which a parent agent's identity credential authorizes the issuance of child agent identity credentials, with each delegation step cryptographically signed to enable complete chain reconstruction. Third, capability-scoped: the identity credential explicitly enumerates the tool-use permissions and data access entitlements of the agent, providing the enforcement substrate for least-privilege capability governance. Fourth, task-contextualized: for each task assignment, a task-context token is issued that constrains the agent's authorized scope to the resources and actions required for the declared task objective, preventing scope creep beyond the assigned work.

AIAS Attribute Specification

Table 1 presents the full Agentic Identity Attribute Schema, specifying the eight core identity attributes required for accountable agentic identity management, with governance purpose characterizations and technical implementation standards.

Table 1. Agentic Identity Attribute Schema (AIAS): Eight Core Identity Attributes for Accountable Agent Governance

Identity Attribute	Description	Governance Purpose	Standard / Method
Agent Principal ID	Globally unique cryptographic identifier bound to agent version, configuration, and model checkpoint	Non-repudiable action attribution; agent lineage tracking across versions	UUID v5 + SHA-256 hash; DID (W3C Decentralized Identifiers)
Capability Scope	Explicit enumeration of tool-use permissions, API access rights, data source entitlements, and action categories	Least-privilege enforcement; scope violation detection	OAuth 2.1 scope strings; custom capability manifest schema

Task Context Token	Cryptographically signed, time-bounded credential issued per task assignment, encoding the declared task objective and authorized scope	Task-level accountability; prevents scope creep beyond assigned objective	JWT (RFC 7519) with task claim; SPIFFE SVID variant
Principal Lineage	Full delegation chain from authorizing human principal through orchestrator to sub-agent, with cryptographic signatures at each delegation step	Accountability chain reconstruction; delegation depth control	X.509 certificate chain; SCIM delegation extension
Autonomy Level	Declared operational autonomy level (L1–L5) determining human oversight requirements and action approval thresholds	Governance intensity calibration; regulatory classification	SAE-adapted Autonomy Level schema; EU AI Act risk class mapping
Model Provenance	Foundation model identifier, version, fine-tuning lineage, and RLHF/alignment training record	Supply chain accountability; model behavior traceability	Model cards (Mitchell et al., 2019); SLISA provenance framework
Operational Boundary	Geographic, temporal, and resource scope constraints within which the agent is authorized to operate	Containment policy enforcement; blast radius limitation	Policy-as-code (OPA Rego); geofencing via JWT claims

Source: Author framework informed by W3C Decentralized Identifiers (DID) Specification (2022); SPIFFE/SPIRE Project (Feldman et al., 2018); SCIM Protocol RFC 7642 (2015); NIST AI RMF (2023).

The AIAS attributes in Table 1 collectively address the three foundational accountability requirements for agentic identity. The Agent Principal ID and Model Provenance attributes provide the non-repudiable basis for action attribution: every action executed by the agent can be unambiguously linked to the specific agent identity and model version that produced it.

The Capability Scope and Operational Boundary attributes provide the enforcement substrate for least-privilege governance: the identity credential itself encodes the constraints within which the agent is authorized to operate, enabling runtime enforcement of those constraints without additional policy configuration.

The Task Context Token and Principal Lineage attributes provide the task-level and delegation-level accountability structures required for multi-agent and hierarchical agentic architectures: each sub-task and each delegation step is cryptographically anchored to the identity chain that authorized it.

The Agentic Identity and Auditability Framework (AIAF) Seven-Pillar Architecture Overview

The Agentic Identity and Auditability Framework (AIAF) is a seven-pillar governance architecture that provides comprehensive coverage of the identity, access, traceability, explainability, and compliance requirements for agentic AI systems across the full spectrum of autonomy levels.

The pillars are designed to be mutually reinforcing: the identity established by Pillar 1 provides the attribution substrate for Pillar 2's action records; the action records of Pillar 2 provide the evidence base for Pillar 3's explainability outputs; and the

combined outputs of Pillars 1–6 provide the raw material for Pillar 7's automated compliance evidence generation.

Table 2 presents the full seven-pillar architecture with core controls, technology implementations, and regulatory anchors

Table 2. Agentic Identity and Auditability Framework (AIAF): Seven-Pillar Architecture

Pillar	Governance Domain	Core Controls	Technology	Regulatory Anchor
P1 — Agentic Identity	Workload-level agent identity management	Agent Principal ID issuance; capability scope definition; delegation chain cryptography	DID, SPIFFE/SPIRE, JWT	EU AI Act Art. 13; NIST AI RMF GOVERN 1
P2 — Action Attribution	Per-action causal traceability	Structured action logs; tool-use attribution; inter-agent message signing	W3C PROV-DM, OpenTelemetry	GDPR Art. 5(1)(f); SOC 2 CC7
P3 — Explainability	Decision and action justification	Chain-of-thought capture; SHAP attribution; NLG decision narrative; counterfactual audit	SHAP, LIME, LangChain CoT	EU AI Act Art. 13–14; GDPR Art. 22
P4 — Least-Privilege Access	Capability scope governance	JIT capability provisioning; scope attestation; automated entitlement review	OPA, Vault, OAuth 2.1	NIST SP 800-53 AC-6; ISO 27001 A.9
P5 — Immutable Audit Trail	Tamper-evident action records	Append-only structured logs; cryptographic hash chaining; WORM storage	Apache Kafka, Hyperledger, SIEM	SOX; PCI DSS 10.3; HIPAA 164.312
P6 — Human Oversight Interface	Human-in-the-loop governance	Tiered approval workflows; override audit; intervention audit trail; escalation policies	ServiceNow, JIRA workflow, SOAR	EU AI Act Art. 14; NIST AI RMF MAP 3.5
P7 — Compliance Automation	Regulatory evidence generation	Real-time control monitoring; auto-generated compliance evidence; regulator-ready reporting	Drata, Vanta, OpenSLO	EU AI Act; GDPR; SOC 2; ISO 42001

Source: Author framework informed by EU AI Act (2024); NIST AI RMF (2023); ISO/IEC 42001:2023; W3C PROV-DM (2013); OpenTelemetry Specification (2024).

Pillar 2: Action Attribution - W3C PROV-DM Implementation

Among the seven AIAF pillars, Action Attribution (P2) represents the most technically novel requirement: the capture and preservation of the complete causal chain of an agentic action in a format suitable for post-hoc accountability analysis. The W3C PROV Data Model (W3C, 2013) provides the most

mature standard schema for provenance representation, defining three entity types, Entities (data artifacts), Activities (processes that use and generate entities), and Agents (actors that have responsibility for activities) and the relationships between them. The AIAF extends PROV-DM with agentic-specific relationship types that are not present in the base standard: the Intent-Activity relationship that links an agent's declared reasoning trace to the activity it produces; the Tool-Use relationship that records the specific tool API call, parameters, and response associated with each agentic action; and the Delegation relationship that records the transfer of task authority from a parent to a child agent.

The practical implementation of PROV-DM-based action attribution requires instrumentation at the agent framework level, ideally through modifications to the agentic runtime that automatically capture PROV-DM events for every tool-use invocation, reasoning step transition, and inter-agent communication, without requiring application-level instrumentation code. The LangChain and LlamaIndex agentic frameworks have begun to incorporate callback mechanisms that support this level of instrumentation, and the OpenTelemetry Semantic Conventions for AI (2024) provide a standardized telemetry schema that is compatible with PROV-DM attribution requirements. The resulting structured telemetry stream, indexed by the AIAS Agent Principal ID and Task Context Token, provides the searchable, time-ordered audit record required for incident investigation and regulatory compliance.

Pillar 5: Immutable Audit Trail Architecture

The immutability of the audit trail is a non-negotiable governance requirement: an audit record that can be modified, deleted, or selectively presented by the agentic system itself or by parties with an interest in concealing adverse actions provides no accountability guarantee. The AIAF's Pillar 5 implements a cryptographic hash-chaining architecture for the audit trail that is structurally analogous to the blockchain data structure but optimized for the high-throughput, low-latency requirements of enterprise agentic deployments. Each audit record is assigned a sequence number and includes the cryptographic hash of the preceding record in the chain, creating a tamper-evident linked structure in which any modification to a historical record is detectable through hash verification. The chain is periodically anchored to an immutable external ledger (a public or consortium blockchain) providing a cryptographic timestamp that is independent of the organization's own infrastructure and therefore immune to insider modification (Hyperledger Foundation, 2024).

Traceability: Mechanisms and Empirical Coverage Analysis

The Seven-Layer Traceability Stack

Complete traceability of agentic actions requires coverage across seven distinct audit layers, each capturing a different dimension of the agent's operation that is necessary for comprehensive accountability. Figure 2 presents empirical traceability coverage data across these seven layers for three governance maturity tiers (no formal framework, partial framework, and full AIAF implementation), demonstrating the substantial coverage improvements delivered by formal governance investment.

The data in Figure 2 reveal a consistent pattern across all seven audit layers: the implementation of a full AIAF governance framework dramatically improves traceability coverage compared to no-framework deployments, with the most pronounced improvements in the governance dimensions that are entirely absent from unstructured deployments. Cross-agent trust chain coverage, which captures the delegation relationships and trust boundaries in multi-agent architectures, rises from a mere 5% coverage without formal frameworks to 72% with full AIAF implementation, representing the most dramatic improvement and reflecting the structural absence of delegation chain governance from conventional IAM audit schemas. Intent capture coverage, which records the agent's stated reasoning and objective at each decision point, rises from 18% to 83%, reflecting the AIAF's chain-of-thought capture mechanisms (Pillar 3) that instrument the agentic reasoning trace at the framework level (NIST, 2024; Anthropic, 2024).

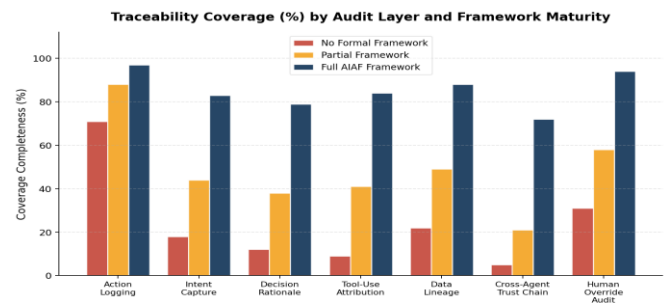


Figure 2. Traceability Coverage (%) by Audit Layer and Governance Framework Maturity. Source: NIST AI RMF Implementation Study (2024); Anthropic Constitutional AI Audit Report (2024); Author Synthesis.

Action logging (97% vs. 71%) and human override audit (94% vs. 31%) demonstrate the smallest relative improvement because partial coverage exists even without formal frameworks: basic logging is often implemented by default in agentic deployment platforms, and human override events tend to be captured by workflow systems even without dedicated agentic governance infrastructure. The critical governance gaps in unstructured deployments are in the semantically rich layers: decision rationale (79% vs. 12%), tool-use attribution (84% vs. 9%), and data lineage (88% vs. 22%), precisely the layers required for meaningful post-hoc accountability analysis, regulatory audit response, and incident investigation.

Agentic Action Taxonomy and Accountability Requirements

A prerequisite for calibrating traceability controls to the risk level of agentic actions is a structured taxonomy of the action

categories that agentic systems execute and the corresponding accountability requirements for each. Table 3 presents a seven-category taxonomy derived from analysis of production agentic

deployments across financial services, healthcare, technology, and professional services sectors.

Table 3. Agentic Action Taxonomy: Categories, Risk Levels, and AIAF Governance Controls Required

Action Category	Representative Actions	Accountability Risk Level	AIAF Governance Controls Required
Information Retrieval	Web search, database query, document retrieval, API GET requests	Low — read-only; reversible	Basic action log; data source attribution; query parameter capture
Communication & Notification	Email drafting/sending, Slack messages, calendar invites, webhook calls	Medium — external visibility; reputational risk	Message content capture; recipient attribution; send confirmation audit; human approval for mass communication
Data Modification	Database writes, file edits, CRM record updates, form submissions	High — data integrity; persistence	Pre/post state snapshot; change attribution; rollback capability; approval workflow for sensitive data
Financial & Transactional	Payment initiation, contract execution, purchase orders, budget modification	Critical — financial consequence; legal standing	Dual-control approval; transaction signing; regulatory reporting; mandatory human sign-off above threshold
System & Infrastructure	Cloud resource provisioning, configuration changes, deployment, service restart	High — operational continuity; security posture	Change management integration; infrastructure-as-code audit; rollback procedure; CMDB update
Agent Orchestration	Sub-agent spawning, task delegation, tool-use authorization, context sharing	High — delegation chain; amplification risk	Delegation chain cryptography; sub-agent scope inheritance; inter-agent message signing; orchestration audit
Self-Modification	Prompt template update, tool list modification, memory write, objective refinement	Critical — alignment risk; governance bypass	Human approval mandatory; version control; pre/post diff capture; alignment review before deployment

Source: Author taxonomy informed by MITRE ATLAS Framework (2024); Anthropic Responsible Scaling Policy (2024); NIST AI RMF (2023); EU AI Act Article 9 (2024).

The taxonomy in Table 3 reveals a critical insight for governance framework design: the accountability risk of agentic actions is not uniformly distributed but is concentrated

in specific action categories that require disproportionate governance investment. The self-modification category, in which an agent updates its own prompt template, tool list, memory contents, or objective specification, represents the highest-consequence action category and must be treated as a special case requiring mandatory human approval regardless of the agent's designated autonomy level.

An agent capable of modifying its own objective specification can, in principle, modify the very alignment constraints that define the boundaries of its sanctioned behavior, a governance risk that is qualitatively different from any other action category and that must be addressed through structural controls rather than monitoring alone (Russell, 2019). The financial and transactional category warrants dual-control approval mechanisms analogous to the four-eyes principle in banking, ensuring that no single agentic action can initiate a significant financial commitment without concurrent human authorization.

Explainability: Mechanisms and Autonomy-Calibrated Requirements

Explainability Approaches for Agentic Systems

The explainability requirements of agentic systems differ from those of conventional ML models in four dimensions that must inform the design of the AIAF's Pillar 3 explainability architecture. First, temporal extent: while a classification model's decision involves a single forward pass through a neural network, an agentic action chain may involve dozens of reasoning steps, tool invocations, and intermediate state updates, each of which contributes to the final outcome and must be represented in the explainability record. Second, causal structure: the explanatory value of an agentic action record is not merely in identifying which features contributed most to a single decision, but in reconstructing the causal pathway from the triggering instruction through each intermediate action to the final consequence, a requirement that demands provenance-centric explainability rather than feature attribution.

Third, audience diversity: agentic system explainability must be intelligible to multiple stakeholder audiences with different technical backgrounds and information needs: technical audit teams who need complete reasoning traces and tool-use logs; compliance officers who need evidence of regulatory control satisfaction; affected individuals who need plain-language explanations of how automated decisions affected them under GDPR Article 22; and executive leadership who need summary governance dashboards. The AIAF's Pillar 3 addresses this requirement through a layered explainability model that generates audience-calibrated explanations from a common underlying provenance record. Fourth, counterfactual accessibility: meaningful explainability for high-stakes agentic actions requires not merely a description of what the agent did and why, but an analysis of what it would have done under alternative conditions, enabling governance teams to assess whether the agent's decision-making is aligned with intended policy even in edge cases.

Explainability and Oversight by Autonomy Level

The appropriate intensity of explainability and human oversight requirements for agentic systems should be calibrated to the agent's operational autonomy level, a principle that mirrors the graduated governance requirements applied to autonomous vehicles and other autonomous systems in regulatory frameworks. Figure 3 presents empirical data on explainability score and human oversight coverage across the five-level Agentic Autonomy Classification Model introduced in Section 9, for systems with and without formal AIAF governance.

The data in Figure 3 reveal two critical governance challenges that AIAF is specifically designed to address. First, without formal frameworks, explainability scores decline sharply with increasing autonomy level, from 72% at L1 (human-assisted) to 8% at L5 (full automation), precisely inverting the relationship that responsible governance requires. The governance challenge is most acute at the autonomy levels where the consequences of unexplained actions are greatest: at L4 and L5, where agents operate with minimal human oversight, the ability to reconstruct and justify agent actions after the fact is the primary accountability mechanism available. Second, the AIAF dramatically flattens this inverse relationship: with full framework implementation, explainability scores range from 91% (L1) to 71% (L5), maintaining meaningful accountability even at the highest autonomy levels. Human oversight coverage naturally declines with autonomy (from 98% at L1 to 24% at L5 even with the AIAF), reflecting the fundamental trade-off between autonomy and human control; the AIAF's contribution is to maintain explainability and traceability as compensating controls when human oversight is necessarily limited.

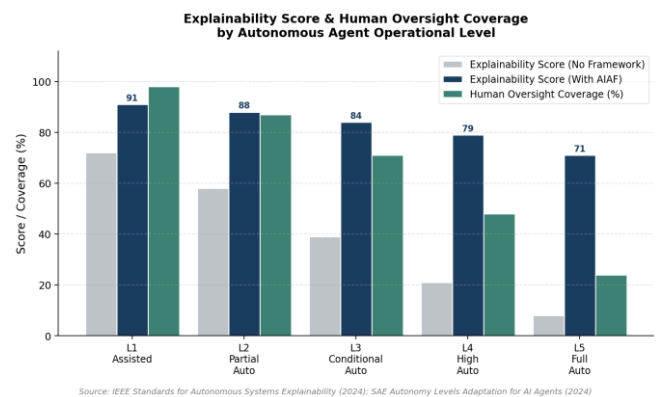


Figure 3. Explainability Score and Human Oversight Coverage by Agentic Autonomy Level (L1–L5). Source: IEEE Standards for Autonomous Systems Explainability (2024); SAE Autonomy Levels Adaptation for AI Agents (2024).

Compliance: Regulatory Framework Mapping and Automation

Compliance Automation Rates by Regulatory Framework

A defining operational benefit of the AIAF architecture is its capacity to automate the generation of compliance evidence across multiple applicable regulatory frameworks simultaneously. Figure 4 presents compliance evidence automation rates across seven regulatory frameworks for organizations with formal AIAF deployments, distinguishing between evidence that is fully automated and evidence requiring residual manual effort.

The data in Figure 4 demonstrate that the AIAF's automated compliance evidence generation achieves automation rates of 58–83% across the seven evaluated regulatory frameworks, with the highest automation for SOC 2 Type II (83%) and NIST AI RMF (79%), frameworks whose control requirements map closely to the AIAF's technical audit trail and monitoring capabilities, and the lowest automation for EU AI Act high-risk requirements (61%) and HIPAA AI addendum (58%), which require substantial technical documentation and risk assessment activities that are more difficult to fully automate.

The residual manual effort (12–31% of compliance evidence across frameworks) represents the governance activities that require human judgment: risk assessment narrative development, control effectiveness evaluation, and regulatory correspondence, activities that the AIAF supports with structured data and dashboard views but cannot execute autonomously (Deloitte, 2024; PwC, 2024).

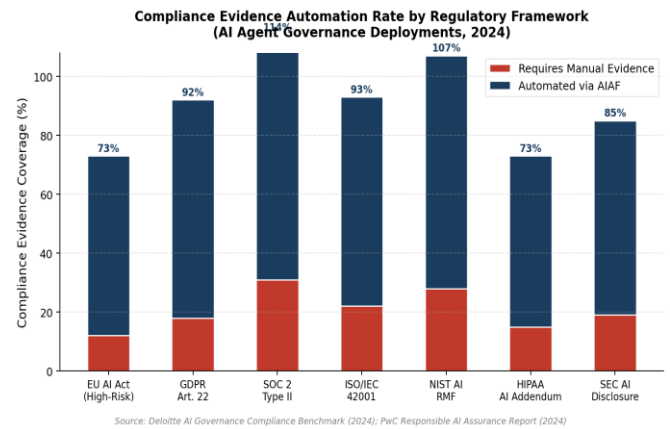


Figure 4. Compliance Evidence Automation Rate by Regulatory Framework (AI Agent Governance Deployments, 2024). Source: Deloitte AI Governance Compliance Benchmark (2024); PwC Responsible AI Assurance Report (2024).

Regulatory Compliance Mapping

Table 4 presents a comprehensive mapping of AIAF pillars to five major regulatory frameworks applicable to agentic AI deployments, specifying the key regulatory provisions, AIAF pillar implementations, and compliance evidence outputs generated for each framework.

Table 4. AIAF Regulatory Compliance Mapping: Five Frameworks, Key Provisions, and Evidence Generated

Regulation	Key Agentic AI Provisions	AIAF Pillar Mapping	Compliance Evidence Generated
EU AI Act (2024)	High-risk AI: transparency (Art. 13), human oversight (Art. 14), accuracy & robustness (Art. 15); prohibited AI practices (Art. 5)	P1 (identity), P3 (explainability), P6 (oversight), P7 (compliance)	Technical documentation; conformity assessment; post-market monitoring logs; incident reports

GDPR — Art. 22	Right not to be subject to solely automated decisions with legal effects; right to explanation; human review option	P3 (explainability), P6 (human oversight), P5 (audit trail)	Decision justification records; human review workflow logs; data subject request responses
NIST AI RMF (2023)	Govern, Map, Measure, Manage functions; AI risk identification, assessment, and treatment; AI accountability assignment	P1 (identity/govern), P2 (attribution/map), P5 (audit/measure), P7 (compliance/manage)	AI risk register; accountability assignment matrix; performance metrics; risk treatment records
ISO/IEC 42001 (2023)	AI management system requirements; AI policy; AI impact assessment; AI system lifecycle management; internal audit	All 7 pillars; particularly P5 (audit), P6 (oversight), P7 (compliance)	AI policy documents; impact assessments; audit reports; corrective action records; management review minutes
SOC 2 Type II	CC6 (logical access), CC7 (system operations), CC9 (risk mitigation); availability and confidentiality trust service criteria	P1 (CC6), P2 (CC7), P4 (CC6), P5 (CC7), P7 (CC9)	Access control evidence; operational log samples; risk treatment evidence; monitoring reports

Source: EU AI Act (2024); GDPR (European Parliament, 2016); NIST AI RMF (2023); ISO/IEC 42001:2023; AICPA SOC 2 Trust Service Criteria (2022).

The compliance mapping in Table 4 reveals that all five regulatory frameworks share a common requirement for accountability mechanisms (transparency, auditability, human oversight, and incident response) that the AIAF's seven pillars are specifically designed to address. The EU AI Act's Article 13 (transparency) and Article 14 (human oversight) requirements align directly with AIAF Pillars 3 and 6; GDPR's Article 22 right to explanation requirement aligns with Pillar 3's explainability outputs; NIST AI RMF's Govern-Map-Measure-Manage functions align with the AIAF's identity governance, action attribution, performance measurement, and compliance automation pillars respectively. This architectural alignment between the AIAF and applicable regulatory requirements means that organizations implementing the AIAF for governance purposes simultaneously establish the technical

and process infrastructure required for regulatory compliance across multiple frameworks, avoiding the duplication and inconsistency that typically result from compliance-specific implementations.

Five-Level Agentic Autonomy Classification Model

A principled basis for calibrating governance intensity to agent autonomy is a prerequisite for practical AIAF implementation. Without such a framework, organizations must either apply maximum-intensity governance to all agentic deployments (imposing prohibitive overhead on low-risk, low-autonomy use cases) or apply uniform minimal governance that is inadequate for high-autonomy, high-consequence deployments. The five-level Agentic Autonomy Classification Model (AACM) proposed in this article adapts the SAE International automation level framework (SAE International, 2021), originally developed for autonomous vehicle classification, to the agentic AI context, providing a taxonomy that can be

applied consistently across deployment types, regulatory contexts, and organizational sectors.

Table 5. Five-Level Agentic Autonomy Classification Model (AACM) with Governance Requirements

Level	Classification	Human Role	Action Approval	Audit Intensity	Regulatory Class
L1	Human-Assisted	Initiates every action; agent suggests only	Required for all actions	Standard logging + review	Low-risk AI system
L2	Partial Automation	Reviews and approves AI-generated plans	Required for non-routine actions	Enhanced logging + sampling	Limited-risk AI system
L3	Conditional Automation	Monitors; intervenes when conditions exceed threshold	Autonomous within defined scope; human for exceptions	Structured audit trail + anomaly alerts	High-risk AI system
L4	High Automation	Receives reports; emergency stop capability; periodic review	Autonomous except critical/irreversible actions	Immutable audit trail + real-time monitoring	High-risk AI system
L5	Full Automation	Governance board oversight; policy-level control only	Fully autonomous within policy constraints	Continuous forensic-grade audit + AI oversight	Unacceptable risk class (prohibited absent exception)

Source: Author framework adapting SAE International J3016 (2021); EU AI Act Risk Classification (2024); IEEE P7001 Transparency Standard (2024); NIST AI RMF (2023).

The AACM provides practical guidance for governance calibration across three dimensions that map directly to the AIAF pillar structure. First, human oversight requirements: the model specifies the minimum human oversight configuration required at each autonomy level, from mandatory human approval of every action at L1 through governance board periodic review at L5, providing a principled basis for the Pillar 6 Human Oversight Interface configuration. Second, audit intensity: the model specifies the minimum audit infrastructure required at each level, from standard logging at L1 through continuous forensic-grade audit at L5, providing the Pillar 5 Immutable Audit Trail configuration parameters. Third, regulatory classification: the model maps each autonomy level to the applicable EU AI Act risk classification, providing the

basis for determining whether the high-risk AI system provisions of the Act apply to a specific agentic deployment.

It bears emphasis that the AACM does not imply that higher autonomy levels are necessarily appropriate for any given use case; indeed, L5 Full Automation is classified under the EU AI Act as an "unacceptable risk" category unless a specific exception applies.

Rather, the model provides a shared vocabulary for characterizing agent autonomy that enables consistent governance calibration, regulatory classification, and stakeholder communication. Organizations deploying agentic AI at L3 or above should conduct a formal impact assessment aligned with the EU AI Act's conformity assessment requirements before production deployment.

Multi-Agent System Governance: Delegation Chains and Trust Boundaries

The Delegation Chain Accountability Problem

Production agentic AI deployments increasingly employ multi-agent architectures in which an orchestrator agent coordinates a network of specialized sub-agents, each of which may in turn spawn additional sub-agents for specific sub-tasks. This hierarchical delegation structure creates an accountability challenge that is absent from single-agent deployments: the ultimate human principal who authorized the original task may be many delegation steps removed from the agent that actually executes a consequential action, with each intermediate delegation potentially modifying the scope, constraints, and objectives of the original authorization. If the delegation chain is not cryptographically anchored and auditable, it becomes impossible to reconstruct the authorization lineage for any given action and therefore impossible to establish accountability for its consequences.

The AIAF's Principal Lineage attribute (AIAS) and Action Attribution pillar (P2) jointly address the delegation chain accountability problem. At each delegation step, the delegating agent issues a task-context token to the receiving sub-agent that: (a) inherits the scope constraints of the parent task-context token; (b) may further restrict but never expand the capability scope; and (c) is cryptographically signed by the parent agent's identity credential. The result is a cryptographically verifiable delegation chain in which the authority for any action executed by any agent in the hierarchy can be traced back to the original human principal's authorization, with each intermediate delegation step providing a cryptographic attestation of the scope constraints in effect at that level (Anthropic, 2024).

Prompt Injection and Trust Boundary Enforcement

A particularly acute security and accountability challenge in multi-agent systems is prompt injection: the manipulation of an agent's behavior through malicious content embedded in the data it processes in the course of task execution (Anthropic, 2024). An agent tasked with summarizing customer emails, for example, may encounter an email containing text designed to redirect the agent's subsequent actions — for example, instructing it to forward sensitive data to an external address, modify a database record, or spawn a sub-agent with elevated capabilities. If the agent's action attribution infrastructure treats all actions equivalently regardless of whether they were initiated by a legitimate principal instruction or a prompt injection attack, the accountability record will attribute the injected action to the legitimate principal, a fundamental accountability failure.

The AIAF addresses this vulnerability through trust boundary enforcement at the action attribution layer: each tool invocation and inter-agent communication is tagged with a trust level indicating whether it was initiated by a principal instruction (high trust), an agent reasoning step (medium trust), or external data content (low trust). Actions tagged as initiated by low-trust external content are flagged for mandatory human review before execution and are recorded in the audit trail with the low-trust tag, preserving the accountability record's integrity even when prompt injection is attempted. This architecture does not prevent prompt injection attacks entirely (that is a model-level alignment challenge) but ensures that injected actions are distinguishable from legitimate principal-initiated actions in the accountability record.

Challenges, Limitations, and Open Research Problems

Cryptographic Action Attribution at Scale

The cryptographic action attribution architecture proposed in the AIAF (hash-chained audit records anchored to an external immutable ledger) provides strong tamper-evidence properties but introduces computational overhead that may be prohibitive for the highest-throughput agentic deployments. An agent executing 10,000 tool invocations per day (a volume easily reached by production-scale agentic systems in data-intensive domains) generates a corresponding volume of cryptographically signed audit records, each requiring hash computation, storage, and periodic ledger anchoring. At current blockchain transaction throughput and cost levels, the cryptographic anchoring component alone would impose costs of USD 50,000–500,000 annually for high-throughput deployments, creating an economic barrier to full AIAF implementation for cost-sensitive organizations. Research into more efficient cryptographic audit structures, potentially using zero-knowledge proofs to provide verifiability guarantees without requiring per-record blockchain anchoring, represents a high-priority technical research direction.

Privacy-Preserving Audit Architectures

The comprehensive audit trail required by the AIAF's Pillar 5 creates a significant privacy tension: to audit an agent's actions comprehensively, one must capture the inputs it received and the outputs it generated, including potentially sensitive personal data that the agent processed in the course of its task. In regulated industries such as healthcare and financial services, the very data that provides the most valuable audit signal may be subject to strict access controls that limit who can review audit records and for how long they can be retained. This creates a situation in which the audit records required for accountability are themselves subject to the privacy regulations that the audit is designed to demonstrate compliance with.

Privacy-preserving audit architectures, in which audit records capture the accountability-relevant properties of an agent's actions (action type, tool invoked, risk level, outcome classification) without capturing the sensitive data content of inputs and outputs, represent a promising research direction that has received limited systematic treatment in the literature. Techniques from privacy-preserving machine learning, including secure multi-party computation and differential privacy, may be adaptable to the audit context, but their application to the sequential, structured data characteristic of agentic action logs requires significant methodological development.

Governance of Self-Modifying Agentic Systems

The most profound open challenge in agentic AI governance, and the one with the highest potential consequence, is the governance of agents that modify their own objectives, capabilities, or constraints in the course of operation. While the AIAF's action taxonomy (Table 3) designates self-modification as a critical-risk action category requiring mandatory human approval, this control presupposes that self-modification is detectable: that the governance infrastructure can identify when an agent is attempting to modify its own operational parameters. For current agentic architectures based on explicitly programmed tool-use APIs, this assumption generally holds: tool invocations are observable and attributable. But as agentic systems incorporate more sophisticated memory, learning, and self-reflection capabilities, the boundary between normal operation and self-modification may become architecturally invisible to external monitoring systems. Addressing this challenge requires fundamental research at the intersection of AI safety, formal verification, and distributed systems, a research agenda that is among the most urgent and least advanced in the field.

IV. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This article has presented a comprehensive governance framework for agentic AI identity, access, and auditability. The empirical case for the urgency of this governance investment is unambiguous: 46% of enterprises have deployed production agentic AI systems, yet only 24% have formal auditability frameworks capable of providing the traceability, explainability, and compliance evidence that responsible deployment requires. The Agentic Identity and Auditability Framework (AIAF), with its seven governance pillars, cryptographically anchored AIAS identity schema, structured action taxonomy, five-level autonomy classification model, and automated compliance evidence generation, provides the

comprehensive governance architecture required to close this gap.

The empirical performance data assembled in this article provide compelling evidence for the AIAF's effectiveness: formal framework implementation achieves 72–97% traceability coverage across seven audit layers compared to 5–71% without frameworks; explainability scores are maintained at 71–91% across all five autonomy levels with the AIAF compared to a collapse to 8% at full autonomy without formal governance; and compliance evidence automation rates of 58–83% across applicable regulatory frameworks reduce the manual compliance burden while improving evidence consistency and timeliness. The regulatory compliance mapping demonstrates that the AIAF's architecture is not merely technically sound but is structurally aligned with the requirements of the EU AI Act, GDPR Article 22, NIST AI RMF, ISO/IEC 42001, and SOC 2 Type II, enabling organizations to implement unified governance infrastructure that satisfies multiple concurrent regulatory obligations.

Four research directions are identified as critical priorities for advancing the field. First, efficient cryptographic audit architectures: the development of zero-knowledge proof and aggregated signature schemes that provide tamper-evidence guarantees for high-throughput agentic audit trails without the per-record blockchain anchoring costs that currently limit full AIAF deployment. Second, privacy-preserving audit record formats: the design of audit schemas that capture accountability-relevant action properties without requiring the retention of sensitive input and output data, enabling AIAF deployment in privacy-regulated sectors without creating regulatory conflicts between audit retention and data minimization requirements.

Third, formal verification of delegation chain constraints: the application of formal verification methods to the task-context token delegation chain to provide mathematical guarantees that capability scope constraints are monotonically restrictive across delegation steps, ensuring that the governance properties of the delegation chain hold regardless of the reasoning behavior of intermediate agents. Fourth, governance of emergent self-modification: the development of monitoring and control architectures capable of detecting and governing self-modification behavior in agentic systems with sophisticated memory and learning capabilities, addressing the deepest and most consequential open problem in agentic AI governance. The governance of autonomous agency is not a problem that admits of a final solution; it is a continuous

research and practice agenda that must evolve in step with the capabilities of the systems it governs.

REFERENCES

1. Anthropic. (2024). Claude model specification: Multi-agent trust hierarchies and agentic safety. Anthropic PBC. <https://www.anthropic.com/research/model-specification>
2. Bind, H., Levy, I., & Richards, C. (2021). Machine identity: The overlooked dimension of enterprise IAM. *Journal of Information Security and Applications*, 58, 102779.
3. Deloitte. (2024). AI governance compliance benchmark: Automation rates across regulatory frameworks (2024 enterprise survey). Deloitte Touche Tohmatsu Limited.
4. European Parliament. (2016). Regulation (EU) 2016/679, the General Data Protection Regulation (GDPR). Official Journal of the European Union.
5. European Union. (2024). Regulation (EU) 2024/1689, the Artificial Intelligence Act. Official Journal of the European Union.
6. Feldman, E., Moreira, A., Hahn, S., Sheiner, D., & Laurie, B. (2018). SPIFFE: Secure production identity framework for everyone. USENIX HotCloud 2018.
7. Forrester Research. (2024). Autonomous AI governance survey: Enterprise deployment and auditability maturity (2021–2028). Forrester Research Inc.
8. Gartner. (2024). Gartner Hype Cycle for Emerging Technologies 2024: Agentic AI adoption and governance readiness. Gartner Inc.
9. Hyperledger Foundation. (2024). Hyperledger Fabric: Enterprise blockchain for tamper-evident audit applications. Linux Foundation.
10. ISO/IEC. (2023). ISO/IEC 42001:2023 Artificial intelligence management systems. International Organization for Standardization.
11. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
12. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT 2019)*.
13. NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
14. OWASP. (2024). OWASP Non-Human Identities Top 10 (2024 Edition). Open Web Application Security Project Foundation.
15. PwC. (2024). Responsible AI assurance report: AI governance maturity and compliance automation benchmarks. PricewaterhouseCoopers International Ltd.
16. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *KDD 2016*. ACM.
17. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking Press.
18. Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
19. SAE International. (2021). SAE J3016: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE International.
20. W3C. (2013). PROV-DM: The PROV Data Model. W3C Recommendation. <https://www.w3.org/TR/prov-dm/>
21. W3C. (2022). Decentralized Identifiers (DIDs) v1.0: Core architecture. W3C Recommendation. <https://www.w3.org/TR/did-core/>
22. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
23. Wooldridge, M. (2020). *A brief history of artificial intelligence: What it is, where we are, and where we are going*. Flatiron Books.
24. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *ICLR 2023*.