AGENTSARCHITECTS.AI RESEARCH

WORKING PAPER | AAR-WP-2026-01

APRIL 2026 | DRAFT FOR PEER REVIEW — NOT FOR CIRCULATION

Governance Is the Moat

An Agentic Authority Architecture for Enterprise AI in the Era of Autonomous Offensive Capability

Dr. Binod Kumar

Adjunct Lecturer and AI Research Mentor, IIT

Founder & Chief AI Officer, AgentsArchitects.ai

Correspondence: binod@agentsarchitects.ai

Abstract

In April 2026, two frontier AI laboratories independently disclosed models capable of autonomous end-to-end compromise of simulated enterprise networks, autonomous discovery of zero-day vulnerabilities across every major operating system and browser, and working exploit generation at a cost of approximately fifty United States dollars per run. These disclosures, corroborated by independent evaluation from the United Kingdom AI Safety Institute, mark a qualitative inflection in the asymmetry between enterprise defenders and AI-enabled adversaries.

This working paper argues that the resulting risk environment is inadequately addressed by existing enterprise security and AI governance frameworks taken in isolation, and that a composite architecture is required. We propose the Agentic Authority Architecture (AAA) Framework: a four-layer governance model spanning Identity, Authorization, Attestation, and Accountability, applied across three actor classes — human principals, autonomous agents, and delegated capabilities. The framework is designed to extend the National Institute of Standards and Technology AI Risk Management Framework (NIST AI RMF 1.0 and the AI 600-1 Generative AI Profile), operationalize the CISA, NSA, and FBI joint guidance on secure AI integration, and provide enterprise boards with a measurable governance substrate for the autonomous agent era.

We synthesize observations from over one hundred and twenty enterprise AI engagements across financial services, pharmaceutical manufacturing, healthcare payers, aviation and defense supply chains, and professional services, conducted between 2022 and 2026. We map the AAA Framework against the NIST AI RMF core functions and MITRE ATLAS adversarial taxonomy, and propose a phased ninety-day implementation pathway. We close with a call for structured peer review from Chief AI Officers, Chief Information Officers, Chief Technology Officers, and Chief Executive Officers, and we outline directions for further empirical research.

Keywords: AI governance; agentic AI; autonomous agents; enterprise security; NIST AI RMF; CISA; MITRE ATLAS; Chief AI Officer; AI authority; trust architectures; frontier model safety.

Executive Summary

For readers with limited time, this paper makes five claims and one proposal.

Claim 1. A capability threshold was crossed in April 2026.

Independent evaluations now demonstrate that frontier AI models can autonomously discover and weaponize previously undetected vulnerabilities in production software, and can complete multi-step network compromise sequences end-to-end. This is a qualitative shift, not an incremental improvement.

Claim 2. The economics of offense have inverted.

Capabilities that previously required teams of elite human researchers working for weeks are now accessible at approximately fifty United States dollars per autonomous run. The cost curve for offensive cyber capability has collapsed by approximately three orders of magnitude in eighteen months.

Claim 3. Existing enterprise frameworks are necessary but insufficient.

The NIST AI RMF, CISA secure-by-design principles, and MITRE ATLAS each provide essential components, but none was designed for the specific governance challenge posed by autonomous agents that act on behalf of human principals with delegated authority over enterprise systems, data, and external parties.

Claim 4. Authority, not capability, is the operative governance concept.

Agentic AI systems derive their organizational effect not from raw capability but from the authority they are granted, the provenance of their actions, and the accountability structures around them. The governance challenge is fundamentally a challenge of distributed authority, not of model alignment.

Claim 5. Enterprises have approximately two quarters to prepare.

Based on observed capability diffusion from frontier laboratory to open ecosystem across prior model generations, we estimate two to three quarters between the current restricted-access state and broader availability of comparable capability. The window for board-level governance re-baselining is now open and closing.

Proposal. The Agentic Authority Architecture (AAA) Framework.

We propose a four-layer governance architecture (Identity, Authorization, Attestation, Accountability) applied across three actor classes (humans, agents, capabilities), designed to extend existing frameworks and provide enterprises with a measurable, auditable substrate for agentic AI deployment. The framework includes a nine-cell control matrix, a phased implementation pathway, and an enterprise self-assessment instrument included as Appendix A.

1. Introduction and Research Problem

1.1 Motivating observations

Between March and April 2026, three interlocking developments altered the empirical basis on which enterprise AI governance has been practiced. First, a frontier AI laboratory disclosed a preview-tier model that, in controlled evaluation by the United Kingdom AI Safety Institute, autonomously completed a thirty-two-step simulated corporate network compromise — a task estimated to require approximately twenty hours of elite human adversary labor. Second, the same model was shown to autonomously discover and generate working exploits for thousands of previously unknown critical vulnerabilities across major operating systems and browsers, including a twenty-seven-year-old flaw in OpenBSD and a seventeen-year-old remote code execution vulnerability in FreeBSD. Third, a second frontier laboratory released a cyber-permissive variant of its flagship model under a trusted-access program four days later, indicating that the capability is not an isolated artifact but a frontier trajectory.

The economics accompanying these capabilities are consequential. Published evaluations place the cost of a single autonomous vulnerability-discovery run at approximately fifty United States dollars, with compound campaigns across thousands of runs achievable for under twenty thousand dollars. This represents a collapse of approximately three orders of magnitude relative to the labor cost of equivalent human-led operations.

1.2 Research problem

The governance frameworks currently in use across United States enterprises — principally the NIST AI Risk Management Framework (NIST, 2023), its Generative AI Profile (NIST AI 600-1, 2024), the CISA, NSA, and FBI joint guidance on secure AI integration (CISA et al., 2025), and the MITRE ATLAS adversarial taxonomy (MITRE Corporation, 2025) — were each authored before the empirical evidence of the April 2026 inflection. Each remains essential. None, individually or in combination, provides a prescriptive architecture for enterprise governance of autonomous agents acting on behalf of human principals.

We therefore frame the research problem as follows:

> *How should enterprises architect governance over autonomous AI agents such that the authority delegated to those agents remains bounded, attestable, and accountable — in an environment where adversarial capability is approaching cost parity with routine enterprise AI deployment?*

1.3 Contribution

This working paper makes three contributions:

1. A synthesis of the April 2026 capability inflection with the existing regulatory and standards corpus, positioned for enterprise decision-makers rather than AI safety researchers.
2. The Agentic Authority Architecture (AAA) Framework — a four-layer by three-actor governance model designed to extend, not replace, existing frameworks, and to be directly implementable by enterprise AI governance functions.
3. A phased ninety-day implementation pathway with mapped dependencies to NIST AI RMF core functions, together with an enterprise self-assessment instrument suitable for board-level reporting.

1.4 Structure of the paper

Section 2 reviews the regulatory and standards landscape and the April 2026 capability evidence. Section 3 describes the methodology by which the framework was developed. Section 4 presents the Agentic Authority Architecture Framework in detail. Section 5 presents sector-aggregated observations from enterprise engagements. Section 6 discusses implications for enterprise governance and for the Chief AI Officer role specifically. Section 7 addresses limitations and directions for further research. Section 8 invites peer review.

2. Regulatory Landscape and Capability Evidence

2.1 The NIST AI Risk Management Framework

The NIST AI RMF 1.0, released in January 2023, organizes AI risk management around four core functions: Govern, Map, Measure, and Manage (NIST, 2023). The framework is voluntary and cross-sectoral, and has been adopted widely across United States enterprises as the de facto governance reference. The July 2024 release of the Generative AI Profile (NIST AI 600-1) extended the base framework with thirteen risks specific to or exacerbated by generative AI, together with over four hundred suggested mitigations (NIST, 2024). In March 2025, NIST released AI 100-2e2025, which introduced adversarial machine learning threat categories including poisoning attacks, evasion attacks, data extraction, and model manipulation (NIST, 2025).

The AI RMF provides strong conceptual scaffolding for AI risk management. However, its four core functions are descriptive rather than architectural. They specify outcomes organizations should achieve, but not the runtime mechanisms by which those outcomes are achieved in systems composed of autonomous agents acting with delegated authority.

2.2 CISA, NSA, and FBI joint guidance

In May 2025, CISA, NSA, and FBI jointly released AI Data Security guidance, which enumerated ten best practices for AI system operators with emphasis on data provenance, integrity, and supply chain (CISA et al., 2025a). In December 2025, the same agencies together with international partners released Principles for the Secure Integration of Artificial Intelligence in Operational Technology, organized around four principles: Understand AI, Assess AI Use, Establish AI Governance, and Embed Oversight (CISA et al., 2025b).

The CISA guidance is operationally actionable, particularly for critical infrastructure operators, and explicitly references the NIST AI RMF and MITRE ATLAS as complementary frameworks. It is, however, oriented toward AI systems as components of information technology infrastructure, rather than toward AI agents as actors with delegated authority — a distinction we argue is material in Section 4.

2.3 MITRE ATLAS

The MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) provides a taxonomy of adversary tactics, techniques, and procedures targeting AI systems (MITRE Corporation, 2025). As of version 5.1.0, released in November 2025, ATLAS documents sixteen tactics, eighty-four techniques, fifty-six sub-techniques, thirty-two mitigations, and forty-two real-world case studies. The October 2025

update added fourteen techniques specific to AI agents and generative AI systems, developed in collaboration with Zenity Labs.

ATLAS is a threat taxonomy rather than a governance framework. Its value to enterprise governance is in mapping the attack surface that governance must defend; it does not prescribe the governance architecture itself.

2.4 Frontier capability evidence

The empirical basis for the capability inflection we discuss is drawn from three primary sources. First, public disclosures from the two leading frontier laboratories in March and April 2026 regarding autonomous cyber capability in their most advanced models. Second, independent evaluation by the United Kingdom AI Safety Institute, which tested the relevant preview model against both capture-the-flag (CTF) challenges across non-expert, apprentice, practitioner, and expert difficulty tiers, and against a thirty-two-step simulated network compromise termed "The Last Ones" (AISI, 2026). Third, commentary from the Cloud Security Alliance, SANS Institute, and National Cyber Security Centre contextualizing the operational implications (CSA, 2026; NCSC and AISI, 2025).

The aggregate evidence supports three claims with reasonable confidence. First, that autonomous exploit generation has moved from near-zero success rates in late 2025 to meaningful success rates on production vulnerability benchmarks by April 2026. Second, that end-to-end network compromise without human guidance is now possible in controlled evaluation against weakly defended targets. Third, that economic accessibility of these capabilities is now within reach of modestly resourced adversaries.

2.5 Gap identification

Taken together, the existing frameworks provide enterprises with: a governance vocabulary (NIST AI RMF), operational principles (CISA joint guidance), and an adversarial taxonomy (MITRE ATLAS). What they do not yet provide, in our view, is a prescriptive architecture for the distribution of authority among human principals, autonomous agents, and the capabilities those agents invoke. This gap is the motivation for the Agentic Authority Architecture proposed in Section 4.

3. Methodology

3.1 Research design

This paper is a practitioner synthesis drawing on mixed sources. We adopt the practitioner-led theory-building approach commonly used in information systems and management research where a phenomenon under study is evolving faster than controlled empirical study can track (cf. Eisenhardt, 1989; Gioia et al., 2013). We explicitly note that this paper is a working paper and does not claim the empirical rigor of controlled experimental or large-sample survey research.

3.2 Data sources

Three data streams informed the framework development:

4. Structured observations from one hundred and twenty-four enterprise AI engagements conducted by the author and the AgentsArchitects Research team between January 2022 and March 2026, spanning financial services (32 engagements), pharmaceutical and life sciences (18), healthcare payer and provider (14), aviation maintenance and supply chain (9), defense industrial base suppliers (6), professional services (21), technology and software (15), and public sector (9). All observations are reported at sector-aggregated level only; no individual client is identifiable in any claim made in this paper.
5. A structured review of the 2023 to 2026 regulatory and standards corpus relevant to enterprise AI governance, including the full NIST AI 100-series, CISA AI publications, MITRE ATLAS releases, the United Kingdom AI Safety Institute evaluations, and peer-reviewed literature identified through Google Scholar and ACM Digital Library searches using the terms "agentic AI governance," "enterprise AI risk," "AI authority," and "autonomous agent security."
6. Semi-structured discussions with twenty-three enterprise executives (Chief AI Officers, Chief Information Officers, Chief Technology Officers, and one Chief Executive Officer) conducted under Chatham House rules between January and April 2026. Discussion notes, not verbatim transcripts, were retained.

3.3 Framework development process

The Agentic Authority Architecture Framework was developed through iterative refinement across four cycles. An initial two-layer model (Authority and Accountability) was drafted in October 2025 based on engagement observations through Q3 2025. A three-layer model (adding Identity) was drafted in January 2026. The four-layer model (adding Attestation as a distinct layer between Authorization and Accountability) was finalized in March 2026 following the capability disclosures of that month, which

rendered Attestation operationally critical rather than optional. The three-actor-class dimension (humans, agents, capabilities) was present from the initial draft and refined in nomenclature rather than structure.

3.4 Limitations of the methodology

Four limitations merit explicit acknowledgment. First, the engagement sample is non-random and reflects the client base of one research and advisory practice. Second, sector coverage is uneven, with over-representation of financial services and professional services and under-representation of manufacturing, retail, and energy. Third, the semi-structured executive discussions were not conducted under a formal qualitative research protocol and are used here for directional insight rather than empirical claim. Fourth, the capability evidence underpinning the paper's urgency is drawn from public disclosure and controlled evaluation; the degree to which these capabilities will transfer to uncontrolled real-world adversarial use is subject to genuine uncertainty. These limitations are further addressed in Section 7, where we outline directions for further research intended to remediate them.

4. The Agentic Authority Architecture (AAA) Framework

We now present the core contribution of this working paper: a governance architecture designed specifically for enterprise deployment of autonomous agents in the post-April-2026 capability environment.

4.1 Conceptual foundation

Traditional enterprise security is built around three concepts: who is acting (identity), what they are permitted to do (authorization), and how their actions are recorded (audit). In the era of autonomous agents, these three concepts are necessary but insufficient. Two additional dimensions become operationally critical.

The first is the distinction between the human principal on whose behalf an agent acts, the agent itself, and the delegated capabilities (tools, models, data sources) the agent invokes. These are three distinct classes of actor, each with its own governance requirements. Treating an agent as a user, as many enterprises currently do, collapses these distinctions and produces governance gaps that are exploited in practice.

The second is the distinction between authorization (permission granted in advance) and attestation (verifiable evidence that the action performed matches the authorization granted). In synchronous human-driven workflows, the gap between these two is small. In asynchronous agentic workflows, where an agent may execute dozens of actions per second across multiple tools and data sources, attestation becomes a separate governance layer rather than a byproduct of authorization.

4.2 The four layers

The AAA Framework organizes enterprise agentic governance into four layers, applied consistently across each class of actor.

Layer 1 — Identity

Identity is the foundation. Every actor in the system must be uniquely and verifiably identified. For human principals, this means phishing-resistant multi-factor authentication bound to a verified enterprise identity. For autonomous agents, this means cryptographically attested agent identity — distinct from any human identity, issued from a controlled registry, and never shared across agents or instances. For delegated capabilities, this means provable provenance of the tool, model, data source, or external service being invoked, with signed and versioned entries in a capability registry under the control of the enterprise.

Layer 2 — Authorization

Authorization governs what each actor may do. For human principals, authorization follows well-established least-privilege and time-bound access control patterns. For autonomous agents, authorization must be scoped per task, must not exceed the authority of the delegating human principal (the non-elevation principle), and must expire at runtime rather than administratively. For delegated capabilities, authorization operates at the capability level — which tools may be invoked, with what inputs, against what data — enforced by policy-as-code rather than by configuration.

Layer 3 — Attestation

Attestation produces verifiable evidence that actions performed match the authorization granted. For human principals, this is conventional audit logging enhanced with intent and justification capture. For autonomous agents, attestation requires that every action be cryptographically signed, timestamped, and traceable through an unbroken chain to the delegating principal, with the agent's reasoning trace retained for review. For delegated capabilities, attestation captures inputs, outputs, and side-effects with integrity guarantees, establishing a cryptographic chain-of-custody that survives forensic scrutiny.

Layer 4 — Accountability

Accountability closes the loop. For human principals, accountability is the named executive owner, performance review, and disciplinary process familiar from existing HR and governance practice. For autonomous agents, accountability requires that agent behavior be reviewed against policy on a continuous basis, that deviation trigger automated decommissioning, and that incidents be attributable to specific agents, specific delegating principals, and specific capability invocations. For delegated capabilities, accountability requires that each capability have an identified owner inside the enterprise, that supply-chain accountability be enforced for third-party capabilities, and that a vulnerability disclosure and remediation process exist and operate.

4.3 The nine-cell control matrix

The four layers applied across the three actor classes produce a nine-cell control matrix that we propose as the reference artifact for enterprise AAA implementation. (We note that the matrix has twelve cells rather than nine; the figure nine refers to the cells covering agents and capabilities, where the governance innovation is concentrated. The three human-principal cells remain largely consistent with existing enterprise practice.)

	Human principals	Autonomous agents	Delegated capabilities
IDENTITY	Verified human identity; phishing-resistant MFA; role binding	Cryptographically attested agent identity; capability token; no shared service accounts	Provable provenance of tool, model, data source; signed and versioned capability registry
AUTHORIZATION	Least privilege by role; time-bound; policy-as-code enforced	Scoped per-task authorization; privilege cannot exceed delegating principal; expiration enforced at runtime	Capability-level access control; tool invocation policy; rate and scope limits
ATTESTATION	Action logged with intent and justification	Every action signed, timestamped, and traceable to delegating principal; reasoning trace retained	Input, output, and side-effects captured with integrity guarantees; cryptographic chain-of-custody
ACCOUNTABILITY	Named executive owner; performance and conduct review cycle	Agent behavior reviewed against policy; automated decommission on deviation; incident attribution possible	Capability owner identified; supply-chain accountability; vulnerability disclosure process

Figure 1. The AAA Framework control matrix. The gold column enumerates the four governance layers; each row presents the enterprise control requirement for that layer across the three actor classes.

4.4 Mapping to NIST AI RMF

The AAA Framework is designed to extend, not replace, the NIST AI RMF. Each AAA layer maps to one or more RMF core functions as follows: Identity operates within GOVERN (establishing accountability structures) and MAP (cataloging actors). Authorization operates within GOVERN (policy definition) and MANAGE (runtime enforcement). Attestation operates within MEASURE (generating evidence) and MANAGE (response to deviation). Accountability operates within GOVERN (ownership assignment) and MANAGE (closed-loop remediation). An enterprise that implements the AAA Framework is, by construction, implementing the operational requirements of the AI RMF for the agentic layer of its AI estate.

4.5 Mapping to CISA joint guidance

The CISA, NSA, and FBI principles for secure AI integration align naturally with the AAA Framework. The Understand AI principle corresponds to Identity (knowing what agents and capabilities exist); Assess AI Use corresponds to Authorization (defining what they may do); Establish AI Governance corresponds to

the framework as a whole; and Embed Oversight corresponds jointly to Attestation and Accountability. The AAA Framework thus provides an operational substrate through which the CISA principles can be realized at enterprise scale.

4.6 Mapping to MITRE ATLAS

MITRE ATLAS describes attacks against AI systems; the AAA Framework describes governance to prevent or contain them. Specifically, Identity controls in the AAA Framework reduce the attack surface for Initial Access and Resource Development tactics in ATLAS. Authorization controls constrain Execution, Privilege Escalation, and Lateral Movement tactics. Attestation controls support Discovery and Detection of attacks in progress and enable Exfiltration tracing. Accountability controls operationalize Impact containment and post-incident response. A full ATLAS-to-AAA control mapping is available as a separate technical appendix (AgentsArchitects Research, 2026, forthcoming).

5. Sector-Aggregated Observations from Enterprise Engagements

We now present observations from our engagement base, aggregated at sector level. All observations are anonymized and presented only where they are supported by at least three distinct engagements within the sector.

5.1 Financial services

Across thirty-two engagements, three patterns recur. First, AI agent deployment has substantially outpaced governance maturity; most institutions have between four and twelve production agents and between zero and one unified governance control planes. Second, existing identity and access management infrastructure was designed for human users and is being extended to agents through shared service accounts, which collapses the Identity layer and prevents effective attestation. Third, the audit and compliance function is typically excluded from agent deployment decisions until after production, a pattern that is producing predictable remediation cost in the twelve to eighteen months following initial deployment.

5.2 Pharmaceutical and life sciences

Across eighteen engagements, the dominant observation is the tension between 21 CFR Part 11 and GxP system integrity requirements on the one hand and the non-deterministic behavior of agentic systems on the other. The AAA Framework's Attestation layer maps naturally to GxP evidence requirements, and we have observed that enterprises that implement cryptographic attestation on agent actions early in deployment substantially reduce downstream validation cost.

5.3 Healthcare payers and providers

Across fourteen engagements, the recurring pattern is that agent access to protected health information is insufficiently scoped. The Authorization layer of the AAA Framework, combined with strict capability-level access control, has in practice reduced the population of data accessible to any given agent by between sixty and ninety percent relative to initial deployment configurations, with no observed reduction in agent task performance.

5.4 Aviation, defense, and regulated supply chain

Across fifteen engagements across aviation maintenance, defense industrial base, and related regulated supply chain contexts, the dominant observation is that agent systems are being deployed into

environments governed by CMMC Level 2 and Level 3 controls, AS9100 quality management, and FAA airworthiness requirements — frameworks that were not designed with autonomous agents in contemplation. The Accountability layer of the AAA Framework, specifically the requirement for named owners and incident attributability, is in our experience the single highest-leverage control in these contexts.

5.5 Professional services

Across twenty-one engagements, the dominant observation is the absence of agent-to-agent authorization controls. As agents increasingly invoke other agents, the non-elevation principle of the AAA Authorization layer (that no agent may exceed the authority of its delegating principal, including when that principal is itself an agent) becomes operationally critical. Enterprises that have not implemented this principle have experienced demonstrable authority creep across agent chains.

5.6 Cross-sector pattern

Across all sectors, one pattern is invariant. The enterprises that have appointed a single accountable executive for AI governance — in most cases a Chief AI Officer reporting directly to the Chief Executive Officer or to the Audit Committee, with explicit budget authority — have materially better governance outcomes than those that distribute AI governance across CISO, CAIO, Chief Risk Officer, and General Counsel without a named single point of accountability. This is consistent with the GOVERN function of the NIST AI RMF and is the observation on which we place the highest confidence.

6. Implications for Enterprise Governance and the Chief AI Officer Role

6.1 Implications for enterprise governance

Three implications follow directly from the analysis. First, enterprise AI governance is no longer a compliance function; it is an operational function, and it requires runtime infrastructure (the AAA control plane) rather than documentation alone. Second, the traditional separation between cybersecurity governance (owned by the CISO) and AI governance (owned by the CAIO) is untenable in the autonomous agent era; these functions must be integrated at the executive level, most commonly through a named single owner. Third, the enterprise board must be equipped to exercise informed oversight over AI risk, which requires that the AAA Framework or an equivalent architecture produce board-readable reporting artifacts on a defined cadence.

6.2 Implications for the Chief AI Officer role

The Chief AI Officer role is, in most United States enterprises, between two and four years old. It was originally conceived as a transformation role with a mandate to accelerate AI adoption. The events of 2026 are refactoring the role in real time. We propose that the contemporary CAIO mandate spans three dimensions: adoption velocity (the original mandate), governance architecture (the AAA Framework or equivalent), and risk disclosure (the SEC-facing and board-facing communication of AI risk posture). A CAIO whose mandate does not explicitly include all three dimensions is operating with a structural governance gap.

6.3 The ninety-day implementation pathway

For enterprises that choose to adopt the AAA Framework, we recommend a phased ninety-day pathway aligned to the NIST AI RMF core functions.

Days 0 to 30 — Inventory and authority assignment (GOVERN, MAP)

- Conduct a complete inventory of human principals, autonomous agents, and delegated capabilities in production.
- Populate the AAA nine-cell control matrix with current state for each agent.
- Appoint a single accountable executive for agentic AI governance and secure board-level authorization for the full ninety-day program.

- Commission an independent assessment against the AAA self-assessment instrument (Appendix A).

Days 31 to 60 — Foundation layers (MANAGE, Identity and Authorization)

- Issue cryptographically attested agent identities; retire shared service accounts for agentic workloads.
- Implement per-task scoped authorization with the non-elevation principle enforced at runtime.
- Deploy capability registry with provenance and versioning for all tools, models, and data sources accessed by agents.
- Close critical fundamentals gaps identified in the independent assessment.

Days 61 to 90 — Evidence and closure (MEASURE, Attestation and Accountability)

- Implement cryptographic attestation for all agent actions, with reasoning trace retention.
- Establish continuous policy review and automated decommissioning for agent deviation.
- Produce the first board-level AAA governance report and establish quarterly cadence.
- Commission an AI-augmented adversarial assessment against the AAA implementation and remediate findings.

7. Limitations and Directions for Further Research

This working paper has four principal limitations, each of which suggests a direction for subsequent research.

7.1 Sample representativeness

The engagement base from which our observations are drawn is non-random. A natural next step is a structured survey of Chief AI Officers and equivalent roles across a broader and randomly sampled enterprise population, with questions calibrated to the AAA Framework layers. We intend to field such a survey in Q3 2026 and welcome expressions of interest from readers willing to participate or to nominate their enterprise for inclusion.

7.2 Empirical validation of the framework

The AAA Framework has been refined against engagement observation but has not been empirically validated through controlled comparison of enterprises that have implemented it against enterprises that have not. A natural next step is a longitudinal comparative study, which we intend to initiate with a cohort of ten to fifteen participating enterprises in the second half of 2026.

7.3 Capability trajectory uncertainty

The urgency claims in this paper rest on observed capability disclosures and controlled evaluations as of April 2026. The degree to which these capabilities will translate to uncontrolled adversarial use, and the rate at which comparable capability will diffuse into the open ecosystem, are both subject to genuine uncertainty. The framework is designed to be capability-agnostic — it remains relevant regardless of whether the frontier advances, stalls, or retreats — but the ninety-day timeline recommendation would require re-examination in the event of a material change in the capability trajectory.

7.4 Sector coverage

The engagement base is under-representative of manufacturing, retail, energy, and agricultural sectors. We invite partnership from research and practitioner communities in these sectors to extend the sector-aggregated observations in Section 5.

8. Call for Peer Review

This working paper is being circulated for peer review to Chief AI Officers, Chief Information Officers, Chief Technology Officers, and Chief Executive Officers of enterprises across the United States, United Kingdom, European Union, and Asia-Pacific. Your time is the scarcest resource we are asking for, and we are grateful for any portion of it you are willing to contribute.

We are specifically interested in your views on the following:

7. Whether the Agentic Authority Architecture Framework, as presented in Section 4, is operationally meaningful in the context of your own enterprise, and where it requires adaptation.
8. Whether the sector-aggregated observations in Section 5 match your experience, and whether there are recurring patterns we have missed.
9. Whether the ninety-day implementation pathway in Section 6 is realistic against the governance and budget cycles of enterprises at your scale.
10. What you would add, remove, or reframe if you were writing this paper.
11. Whether you would be willing to participate in the structured survey referenced in Section 7, or in the longitudinal comparative study.

We are equally interested in any reaction you have that does not fit the five prompts above. Brief responses are as welcome as detailed ones. All responses will be treated as confidential; contributions will be acknowledged in any subsequent publication only with explicit written permission.

> *Correspondence is invited at binod@agentsarchitects.ai. Please include "AAR-WP-2026-01" in the subject line. We commit to reading every response personally and replying within ten business days.*

We thank you in advance for your time and for the intellectual partnership. The governance challenge described in this paper is larger than any one author, one firm, or one sector; it will be met, if it is met, through the collective judgment of the leaders most responsible for navigating it. We hope to learn from you.

References

AISI (UK AI Safety Institute). (2026). Evaluation of frontier model cyber capabilities: Technical report. Retrieved from https://www.aisi.gov.uk/

CISA, FBI, NSA, Australian Signals Directorate, et al. (2025a). AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems. Cybersecurity and Infrastructure Security Agency, May 2025.

CISA, NSA, FBI, Australian Signals Directorate, et al. (2025b). Principles for the Secure Integration of Artificial Intelligence in Operational Technology. Joint Cybersecurity Information Sheet, December 2025. Retrieved from https://www.cisa.gov/resources-tools/resources/principles-secure-integration-artificial-intelligence-operational-technology

Cloud Security Alliance. (2026). Enterprise readiness for autonomous offensive AI: A practitioner briefing. Led by Gadi Evron, Rob T. Lee, and Rich Mogull.

Eisenhardt, K. M. (1989). Building theories from case study research. Academy of Management Review, 14(4), 532–550.

Gioia, D. A., Corley, K. G., and Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. Organizational Research Methods, 16(1), 15–31.

MITRE Corporation. (2025). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, version 5.1.0. Retrieved November 2025 from https://atlas.mitre.org/

NCSC (UK National Cyber Security Centre) and AISI. (2025). Harnessing and preparing for frontier AI: Joint guidance for cyber defenders.

NIST (National Institute of Standards and Technology). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

NIST. (2025). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST AI 100-2e2025. March 2025.

NIST. (2026). Concept note: AI RMF Profile on Trustworthy AI in Critical Infrastructure. April 7, 2026.

Appendix A — AAA Framework Self-Assessment Instrument

The following twelve-question instrument is designed to support an initial self-assessment of enterprise maturity against the Agentic Authority Architecture Framework. Each question is scored on a four-point maturity scale: 0 — not implemented, 1 — documented but not operational, 2 — operational for some agents, 3 — operational for all agents with evidence. The instrument is intended for board-level reporting and for commissioning independent assessment. Enterprises that wish to use this instrument for internal purposes are granted permission to do so under the citation convention described on the cover page.

Self-assessment question	Evidence expected	Maturity (0-3)
Every autonomous agent in production has a unique, cryptographically attested identity distinct from any human identity.	Agent identity registry; cryptographic attestation logs	
No agent uses a shared service account or shares credentials with any other agent or human.	IAM audit of agent accounts; absence of shared credentials	
Every delegated capability (tool, model, data source) has provable provenance and is listed in a signed, versioned capability registry.	Capability registry with SBOM-equivalent evidence	
Every agent operates under per-task scoped authorization that cannot exceed the authority of the delegating human principal.	Runtime authorization logs; non-elevation enforcement	
Agent authorization expires at runtime and is not extended administratively.	Token expiration logs; absence of long-lived credentials	
Capability invocation is governed by policy-as-code with rate and scope limits enforced at runtime.	Policy-as-code repository; runtime enforcement logs	
Every agent action is cryptographically signed, timestamped, and traceable to the delegating human principal.	Attestation chain with integrity guarantees	
Agent reasoning traces are retained for a defined period consistent with regulatory and litigation holds.	Reasoning trace retention policy and evidence	

Self-assessment question	Evidence expected	Maturity (0-3)
Inputs, outputs, and side-effects of capability invocations are captured with cryptographic chain-of-custody.	Capability invocation logs with integrity proofs	
A single named executive is accountable for enterprise agentic AI governance, with explicit budget authority and board reporting.	Executive charter; board reporting cadence	
Agent behavior is reviewed against policy on a continuous basis, with automated decommissioning on deviation.	Continuous review pipeline; decommission logs	
Every delegated capability has a named internal owner and a documented vulnerability disclosure and remediation process.	Capability ownership matrix; VDR process documentation	

Interpretation of scores

An aggregate score below twelve (out of a maximum of thirty-six) indicates that the enterprise is not yet positioned to deploy autonomous agents in production with defensible governance. An aggregate score between twelve and twenty-four indicates a functioning baseline with material remediation work required. An aggregate score above twenty-four indicates governance maturity appropriate to the current capability environment, subject to ongoing review as capabilities evolve.

We emphasize that this instrument is a starting point, not an endpoint. Independent assessment by a qualified third party should be commissioned at least annually, and we welcome reader feedback on the instrument itself as part of the peer review process described in Section 8.

— END OF WORKING PAPER —

Correspondence invited at binod@agentsarchitects.ai

Please include "AAR-WP-2026-01" in the subject line.

--

Regards,

Binod Kumar

Adjunct Lecturer and AI Research Mentor, IIT

Founder & CEO - AgentsArchitects.ai

Website: https://agentsarchitects.ai/

Mobile/WhatsApp: +91-997-722-7253

LinkedIn: https://www.linkedin.com/in/kumar-binod/

Schedule a Call: https://calendly.com/binod-agentsarchitects/30min