The OSCAR PUBLISHING Services
American Journal of Applied Science and Technology

VOLUME Vol.05 Issue 10 2025
PAGE NO. 276-283

An Intent-Aware Zero-Trust Identity Architecture For Secure Agentic AI In Untrusted Networks

Dr. Elena M. Rossi
Institute for Distributed Systems, University of Zurich, Switzerland

Received: 01 October 2025; **Accepted:** 15 October 2025; **Published:** 31 October 2025

Abstract: Background: The rise of agentic artificial intelligence (AI) — autonomous, goal-driven software entities that act on behalf of users or organizations — introduces novel identity, access, and trust challenges for modern networks. Traditional perimeter-based models of security are ill-suited for environments where autonomous agents, dynamic workloads, and decentralized identities interact across hybrid cloud, on-premises, and edge infrastructures (Gilman & Barth, 2017; Department of Defense CIO, 2007). Recent proposals emphasize integrating Zero Trust principles with intent-aware identity management to protect AI workloads and agentic behaviors (Hasan, 2024; Achanta, 2025; Kumar, 2023).

Objective: This research article proposes a comprehensive, publication-ready architecture — an Intent-Aware Zero-Trust Identity Architecture (IAZTIA) — that unifies human and machine access, supports agentic AI, and enforces continuous, intent-based policy decisions while accounting for non-stationarity, noisy labels, and adversarial behaviors in telemetry and identity signals (Anderson & McGrew, 2017).

Methods: The architecture synthesizes established standards and operational practices including FIPS 199 security categorization, Cloud Security Alliance Secure Device Posture and SDP concepts, hardware asset management, SPIFFE/SPIRE identity federation mechanisms, decentralized identifiers (DIDs), and intent-based network virtualization principles (NIST FIPS 199, 2004; CSA-SDP, 2015; HWAM, 2015; CNCF/SPIFFE, 2024; W3C, 2023; IBNVN, 2013). We describe a layered methodology: identity provenance and binding, intent extraction and semantic normalization, continuous policy evaluation under Zero Trust, telemetry validation and robust learning for noisy labels, and governance controls for accountability and audit. Design choices are grounded in threat and risk taxonomies developed for agentic AI (OWASP, 2024; OWASP Agent Risk, 2024; Syros et al., 2025).

Results: The IAZTIA design presents: (1) identity constructs that bind human, device, and agentic AI identities using short-lived cryptographic credentials and verifiable DIDs; (2) an intent model capturing goals, constraints, and permitted action templates for agents; (3) a policy decision and enforcement fabric leveraging SPIFFE/SPIRE and SDP-aligned micro-segmentation; (4) robust telemetry pipelines applying practices from malware traffic and noisy label research to maintain policy fidelity (Anderson & McGrew, 2017); and (5) governance controls for role separation, lifecycle management, and incident forensics (Hassan, 2025; Bhushan et al., 2025). We further provide attack scenarios and mitigations, and propose measurable metrics for resilience and trustworthiness.

Conclusions: IAZTIA advances the state of practice by explicitly combining intent semantics with Zero Trust identity controls for agentic AI, enabling continuous, contextual access decisions while providing auditability and governance. The architecture addresses known challenges — identity sprawl, telemetry poisoning, credential misuse, and non-stationary behavior of agents — and outlines a path for operational adoption integrating standards and cloud-native identity primitives (Cohen et al., 2013; Gilman & Barth, 2017; W3C, 2023).

Keywords: Agentic AI, Zero Trust, Intent-Aware Identity, Decentralized Identifiers, SPIFFE/SPIRE, Telemetry

Robustness.

INTRODUCTION:

The contemporary cybersecurity landscape is shifting from perimeter-centric defenses to models that assume network compromise and emphasize continuous verification and least privilege. Zero Trust, as an operational paradigm, prescribes that no user or system is inherently trusted based solely on location or previous authentication; instead, every access decision is contextual, continuously evaluated, and minimally privileged (Gilman & Barth, 2017). Simultaneously, the emergence of agentic AI — systems with autonomous decision-making capabilities that perform tasks with varying degrees of human oversight — challenges traditional identity and access management (IAM) because agents act as independent principals, can move across environments, and may modify or create artifacts in subordinate systems (Kumar, 2023; OWASP Agentic AI Security Navigator, 2024).

Agentic AI elevates identity concerns in two critical ways. First, agents need identities that represent their provenance, capabilities, authorized intents, and constraints — an identity model that cannot be reduced to static service accounts or long-lived API keys. Second, the behavior of agentic systems is intrinsically non-stationary and contextual, meaning telemetry and behavioral signals used for authentication and authorization can change over time, and may be noisy or manipulated (Anderson & McGrew, 2017). These characteristics necessitate identity systems that are dynamic, provenance-aware, and intent-sensitive while operating under Zero Trust controls.

Existing literature and operational guidance provide pieces of the solution. Standards and frameworks such as FIPS 199 recommend categorizing information and systems by impact to inform protection needs (NIST FIPS 199, 2004). The Cloud Security Alliance's SDP specification and similar secure access patterns emphasize decoupling network access from location, enabling application-level policy enforcement (CSA-SDP, 2015). SPIFFE and SPIRE define practical, cloud-native identity primitives for workloads that issue short-lived identities across diverse environments (CNCF, 2024). Decentralized Identifiers (DIDs) provide mechanisms for verifiable identities that are not bound to a single provider and enable cryptographic proofs of control (W3C, 2023). Research in intent-based network virtualization shows that intent can be formalized and compiled into network policies (Cohen et al., 2013). Recent security research explicitly addresses identity and governance for agentic AI, advocating intent-aware identity controls and new threat taxonomies (Hasan, 2024; Achanta, 2025; Syros et al., 2025; Hassan, 2025).

Despite this progress, a comprehensive architecture that unifies intent semantics, continuous Zero Trust enforcement, robust handling of noisy telemetry, and governance for agentic AI identities remains absent or fragmented. Many operational deployments still rely on static identities, long-lived credentials, or human-centric access models that do not scale to thousands of autonomous agents operating across cloud, edge, and partner infrastructures (Department of Defense CIO, 2007; HWAM, 2015). The literature documents that telemetry and label noise can dramatically degrade classifiers and monitoring systems — a critical risk when access decisions are based on behavioral signals — requiring robust learning and validation methods (Anderson & McGrew, 2017).

This article addresses the gap by proposing an Intent-Aware Zero-Trust Identity Architecture (IAZTIA) that combines existing standards and pragmatic mechanisms into a coherent, deployable blueprint. IAZTIA synthesizes cryptographic identity binding, intent modeling, continuous policy evaluation, telemetry robustness, and governance controls. It explicitly targets the protection of agentic AI workloads and their interactions with human users, devices, and network services. The remainder of the article details the methodology for designing IAZTIA, presents the architecture and its components, analyzes attack scenarios and mitigations, discusses limitations and operational considerations, and outlines future research and standardization directions drawn from recent works (OWASP, 2024; Achanta, 2025; Kumar, 2023; Syros et al., 2025).

METHODOLOGY

The architecture methodology is intentionally multidisciplinary, combining principles from cybersecurity standards, cloud-native identity mechanics, intent representation, robust telemetry processing, and governance frameworks. The methodological objectives are: (1) to produce identity bindings that are cryptographically provable and short-lived; (2) to capture agent intent in a machine-understandable way that supports fine-grained, contextual access control; (3) to implement continuous evaluation and enforcement consistent with Zero Trust; (4) to design telemetry and learning systems resilient to label noise and adversarial

manipulation; and (5) to embed governance, audit, and lifecycle operations.

Design Rationale and Principles

1. Least Privilege and Continuous Verification: In line with Zero Trust principles, access decisions should grant the minimum permissions needed and continuously re-evaluate them based on context (Gilman & Barth, 2017). This requires short-lived credentials, ephemeral session tokens, and dynamic policy enforcement.

2. Provenance and Verifiability: Agent and device identities must be provably bound to their creators, configurations, and attestations. Decentralized identity primitives (DIDs) and workload identity frameworks (SPIFFE) provide mechanisms for cryptographic verification and cross-domain portability (W3C, 2023; CNCF, 2024).

3. Intent as a First-Class Attribute: Agents operate to achieve goals. Encoding intent — including high-level goals, constraints, and allowed action patterns — allows policies to assess whether an observed action aligns with authorized intent, reducing false positives and enabling semantic policy enforcement (Cohen et al., 2013; Hasan, 2024).

4. Robust Telemetry and Learning: Telemetry used for authentication and anomaly detection is subject to noise and adversarial manipulation. Methods that explicitly account for noisy labels and non-stationary distributions are necessary to maintain policy fidelity (Anderson & McGrew, 2017).

5. Segmentation and Minimal Blast Radius: Micro-segmentation and SDP-like access patterns reduce lateral movement and exposure, constraining an agent's authorized surface to its necessary resources (CSA-SDP, 2015).

6. Governance, Accountability, and Auditability: Identity lifecycle management, intent signing, and immutable logs support accountability and incident forensics, addressing compliance and audit requirements (Hassan, 2025; Bhushan et al., 2025).

Methodological Components

1. Identity Fabric Design: Define identity types (human, device, workload/agent, service), identity attributes (provenance, creator, version, allowed intents), credential lifecycles (issuance, renewal, revocation), and binding mechanisms. Utilize SPIFFE/SPIRE for workload identities, DIDs for agent identity portability, and hardware attestation for binding to devices (CNCF, 2024; W3C, 2023; HWAM, 2015).

2. Intent Modeling and Normalization: Develop an intent ontology capturing goal types, permissions, constraints, and action templates. Implement semantic normalization that maps heterogeneous intent specifications into a canonical representation consumed by the policy engine (Cohen et al., 2013; Hasan, 2024).

3. Policy Decision and Enforcement Plane: Implement a PDP (policy decision point) and PEP (policy enforcement point) fabric that integrates identity attributes, intent assertions, contextual telemetry, and risk scoring. Policies are expressed in a high-level, intent-aware language and compiled into enforcement rules, using micro-segmentation, API gateway rules, and token scoping to enforce decisions (CSA-SDP, 2015; Gilman & Barth, 2017).

4. Telemetry Pipeline and Robust Learning: Design telemetry ingestion and validation pipelines that perform provenance checks, sanitize inputs, and utilize robust learning models resilient to noisy labels. Where classifiers are used for anomaly detection, incorporate techniques for label noise estimation, model retraining safeguards, and human-in-the-loop verification for high-impact decisions (Anderson & McGrew, 2017).

5. Governance and Lifecycle Controls: Define processes for agent onboarding, intent approval, credential issuance, revocation, and incident response. Incorporate immutable audit logs, signed intent documents, and role-based separation of duties to maintain accountability (Hassan, 2025; Bhushan et al., 2025).

6. Threat Modeling and Risk Assessment: Employ agent-specific threat taxonomies to identify likely attack vectors (e.g., identity spoofing, intent misrepresentation, telemetry poisoning) and map mitigations to architecture components (OWASP Agent Risk, 2024; OWASP AI Threat Modeling, 2024).

Methodological Validation Strategy

The architecture is validated conceptually via scenario-based analysis: representative agentic AI use cases (autonomous orchestration agents, data processing pipelines, and decision support assistants) are exercised against threat scenarios. For each scenario we examine identity issuance, intent propagation, policy evaluation, telemetry anomalies, and governance actions. The validation emphasizes soundness of design, alignment with standards, and coverage of identified threat vectors. Where empirical evaluation would be appropriate (e.g., performance of robust learning techniques), the article specifies experimental protocols and metrics to enable future empirical work.

RESULTS

The proposed IAZTIA is presented as a layered architecture, with detailed descriptions of each component and its interactions. The results section describes the operational behavior of the architecture under typical and adversarial scenarios and provides recommended configurations and metrics for monitoring.

Architectural Overview

IAZTIA organizes into five interacting layers: Identity Fabric, Intent Layer, Policy and Enforcement Plane, Telemetry and Learning Layer, and Governance & Lifecycle Services. Each layer fulfills specific responsibilities and interfaces with others via well-defined artifacts: signed identity assertions, normalized intent statements, policy decisions, telemetry events, and audit logs.

1. Identity Fabric

- Identity Types and Attributes: The fabric recognizes four principal identity types: Human Principals, Device Principals, Workload/Agent Principals, and Service Principals. Each identity carries attributes including issuer provenance, cryptographic keys, role assertions, allowed intent classes, and lifecycle metadata. Binding to hardware identity tokens is encouraged for devices (HWAM, 2015).
- Credential Management: Short-lived X.509 or SPIFFE SVIDs (Secure Identity Documents) are issued to workloads; for agentic AI, identities include signed intent manifests and verifiable DIDs allowing cross-domain portability (CNCF, 2024; W3C, 2023). The fabric supports immediate revocation via token blacklists and attestation revocation lists.
- Attestation and Hardware Roots: Hardware attestation, using TPM or equivalent, binds the runtime instance to a specific image or configuration, supporting non-repudiation and reducing the risk of identity theft (HWAM, 2015).

2. Intent Layer

- Intent Ontology: Intent documents express goals at multiple granularities: Declarative Goals (e.g., “optimize supply chain latency”), Procedural Constraints (e.g., “do not access PII”), and Action Templates (e.g., allowed API calls and resource types). Intent manifests are cryptographically signed by authorized creators and time-bounded.
- Normalization and Semantics: Intent normalization maps provider-specific intent syntaxes into a canonical schema that the PDP consumes. The normalization handles ambiguity by requiring intent owners to specify constraints and risk tolerance levels (Cohen et al., 2013).
- Intent Binding: Each agent identity includes a pointer or embedded signed intent manifest; intent binding is evaluated at issuance and revalidated at renewal, and deviations trigger immediate policy checks.

3. Policy Decision and Enforcement Plane

- PDP and PEP Topology: The PDP consumes identity attributes, intent assertions, telemetry context, and external risk signals to produce allow/deny and policy scope decisions. PEPs (API gateways, service mesh sidecars, and host-level enforcement agents) apply decisions in real time.
- Policy Language and Compilation: Policies are expressed in a high-level, intent-aware language and compiled into actionable rules—e.g., token scopes, network micro-segmentation rules, rate limits, and API filters. Policies incorporate temporal constraints and intent alignment checks (CSA-SDP, 2015).
- Risk Scoring and Conditional Access: A risk engine aggregates telemetry anomalies, provenance signals, and historical behavior to produce risk scores that can dynamically tighten or relax policy scopes.

4. Telemetry and Learning Layer

- Provenance and Sanitization: Telemetry ingestion includes provenance verification (signed events, chained attestations) and sanitization to limit the impact of malformed or adversarial inputs (Anderson & McGrew, 2017).
- Robust Detection Models: For anomaly detection, models are trained with techniques that estimate and correct for noisy labels, use ensemble methods to reduce brittle decisions, and employ windowed retraining to adapt to non-stationary behavior (Anderson & McGrew, 2017).
- Human-in-the-Loop for High-Risk Decisions: For actions with high potential impact, automated decisions are augmented with human review or multi-agent consensus checks to reduce false positives/negatives and to audit intent deviations.

5. Governance & Lifecycle Services

- Onboarding and Approval: Agent onboarding requires intent approval workflows, signature of intent manifests, and assignment of lifecycles and scopes. Role separation ensures different actors (developers, approvers, operators) have distinct responsibilities (Bhushan et al., 2025).
- Audit and Forensics: Immutable logs record identity issuance, intent signatures, policy decisions, and enforcement outcomes. Logs are structured to support reproducible forensics and compliance reporting (Hassan, 2025).

- Revocation and Recovery: Credential revocation, intent revocation, and emergency kill switches are designed to handle compromised agents or runaway behaviors. The governance model includes escalation and containment playbooks.

Operational Scenarios and Analysis

To illustrate architecture behavior, consider three representative scenarios that exercise different parts of IAZTIA.

Scenario A: Autonomous Data Orchestration Agent

An orchestration agent is created to transfer datasets between cloud storage and an analytics cluster. Its intent manifest declares allowed storage buckets, data retention constraints, and prohibition on transferring PII outside approved regions. Upon onboarding, the PDP issues a short-lived SPIFFE SVID bound to the agent and its signed intent manifest. The PEP enforces API scopes, and telemetry monitors for unexpected resource access or deviations in transfer size patterns. If telemetry indicates anomalous access to unapproved buckets, the risk engine increments the risk score, causing the PDP to narrow token scopes or pause the agent for manual review. All events and decisions are logged for audit (CNCF, 2024; Hasan, 2024).

Scenario B: Distributed Decision Assistant in Partner Environments

An assistant agent operates across partner clouds using a DID for identity portability. The intent manifest authorizes read-only queries to shared datasets and prohibits any write operations to partner systems. Policy enforcement is implemented at partner PEPs through mutually recognized intent signatures and short-lived credentials. If a partner's telemetry pipeline reports conflicting provenance attestation (e.g., signature mismatch), the PDP of the requesting domain denies the operation and notifies governance for cross-domain clarification (W3C, 2023; Gilman & Barth, 2017).

Scenario C: Adversarial Telemetry Poisoning Attempt

An adversary attempts to poison anomaly detectors by injecting spurious telemetry to normalize malicious behaviors. The telemetry pipeline's provenance checks and label noise estimation detect inconsistencies and flag suspicious inputs. The robust learning models discount suspected poisoned labels and rely on ensemble consensus; when confidence falls below thresholds, human-in-the-loop review is triggered, and affected agent privileges are reduced preemptively. The attack is contained while false positives are minimized through conservative retraining safeguards (Anderson & McGrew, 2017).

Metrics and Measurable Outcomes

IAZTIA proposes measurable metrics to evaluate effectiveness:

- Mean Time to Detect (MTTD) Intent Deviation: Time between intent deviation occurrence and detection by the telemetry/risk engine.
- Mean Time to Revoke (MTTRv): Time to revoke or reduce privileges following confirmed compromise.
- False Positive/Negative Rates in Anomaly Detection: Evaluated after label noise mitigation and robust learning measures.
- Credential Lifetime and Renewal Frequency: Average lifetime of issued credentials, balancing usability and security.
- Cross-Domain Intent Verification Success Rate: Percentage of intent assertions accepted by partner domains using DIDs and signed manifests.
- Audit Completeness: Coverage of events captured for forensic analysis.

Each metric maps to specific components and provides operational levers: e.g., reducing credential lifetimes shortens blast radius but increases token churn; improving provenance checks reduces false positives but may introduce latency.

Attack Surface and Mitigations

IAZTIA identifies major attack surfaces and prescribes layered mitigations:

- Identity Spoofing: Mitigated by short-lived credentials, hardware attestation, and DID-based verification (W3C, 2023; HWAM, 2015).
- Intent Misrepresentation: Mitigated using signed intent manifests, multi-party approvals for high-risk intents, and audit trails (Hasan, 2024; Bhushan et al., 2025).
- Telemetry Poisoning: Mitigated via provenance checks, robust learning models, ensemble detection, and human review for critical actions (Anderson & McGrew, 2017).
- Token Theft and Replay: Mitigated by SVID short lifetimes, one-time token binding, and cryptographic session binding (CNCF, 2024).
- Lateral Movement: Mitigated through micro-segmentation and SDP principles limiting an agent’s network footprint (CSA-SDP, 2015).
- Supply-Chain Compromise: Addressed by provenance attestation, signed images, and policy checks on runtime configurations (Department of Defense CIO, 2007).

Collectively, these mitigations combine to reduce the

likelihood of undetected high-impact events and to improve containment when events occur.

DISCUSSION

The IAZTIA architecture advances the theoretical and practical discourse on securing agentic AI by treating intent as a central identity attribute and by integrating rigorous Zero Trust mechanisms with robust telemetry and governance. The discussion explores theoretical implications, potential counterarguments, limitations, deployment considerations, and avenues for future research.

Theoretical Implications

1. Reframing Identity Beyond Authentication: Traditional IAM focuses on "who" and "what" (user or service), but IAZTIA elevates "why" — the intent — as a first-class element that informs policy decisions. Intent as an identity attribute bridges the semantic gap between authorization and behavioral expectations. This reframing aligns with intent-based networking principles and extends them to identity control, enabling policy reasoning that is semantically richer and better aligned with business objectives (Cohen et al., 2013).

2. Continuous Contextual Authorization: By combining short-lived cryptographic identities with continuous telemetry and risk scoring, IAZTIA operationalizes the Zero Trust principle of continuous verification. It recognizes identity as a dynamic property contingent on the current context and intent alignment (Gilman & Barth, 2017).

3. Robustness in Non-Stationary Environments: Agentic AI introduces adaptive behaviors that change over time. The incorporation of robust learning techniques acknowledges that telemetry and behavior classifiers face class-imbalance, concept drift, and noisy labels. This is a practical recognition that static models cannot reliably inform critical access decisions in evolving environments (Anderson & McGrew, 2017).

Practical Implications and Counterarguments

1. Operational Complexity: Critics may argue that intent signing, short-lived credentials, and continuous evaluation greatly increase operational overhead. Indeed, IAZTIA requires investment in identity infrastructure (SPIFFE/SPIRE), telemetry pipelines, and governance processes. However, the architecture mitigates operational costs through automation: intent templates, intent normalization, and policy compilation reduce human burden; short-lived credentials can be managed by established identity services; and telemetry provenance can be automated via attestation and signed events (CNCF, 2024; CSA-SDP, 2015).

2. Human-Machine Coordination: Requiring human approval for certain intents may slow adoption, especially for agents designed to act autonomously. IAZTIA addresses this by enabling tiered intent approval: low-risk, high-frequency intents can be auto-approved under guardrails, while high-risk intents require explicit human approval with appropriate auditability (Bhushan et al., 2025).

3. Privacy and Data Protection Concerns: The architecture requires capturing extensive telemetry and identity metadata. Privacy safeguards must be embedded: telemetry minimization, pseudonymization, and purpose-limited logging are necessary to align with data protection principles and to reduce exposure of sensitive data in audit logs (Hassan, 2025).

Limitations and Open Questions

1. Standardization of Intent Ontologies: A core challenge is the lack of widely accepted, interoperable intent ontologies. While IAZTIA defines a canonical schema, broad adoption requires community standards and mappings across domains. Efforts such as intent specifications in network management provide a starting point, but domain-specific ontologies (healthcare, finance, industrial control) will be required (Cohen et al., 2013).

2. Scalability and Performance: Continuous evaluation of thousands of agents with complex intent checks and telemetry processing may introduce latency. Architectural choices (edge PDPs, hierarchical policy caching, and efficient intent compilation) can mitigate performance issues, but empirical evaluation is necessary to quantify trade-offs.

3. Robustness Against Highly Sophisticated Adversaries: While the architecture addresses many attack vectors, nation-state level adversaries or highly motivated attackers with insider access to provenance signing keys pose significant challenges. Hardware attestation and supply-chain controls raise the bar, but absolute guarantees are unattainable; instead, the architecture improves resilience and reduces risk exposure (Department of Defense CIO, 2007).

4. Economic and Organizational Barriers: Adoption implies organizational changes — IAM, DevOps, and security teams must collaborate on intent templates, approval workflows, and governance. Change management and clear ROI arguments are necessary to motivate adoption.

Deployment Considerations

1. Incremental Adoption Path: IAZTIA supports incremental deployment. Early adopters can begin by

issuing signed intent manifests for critical agents, adopting short-lived credentials for sensitive workloads, and implementing telemetry provenance checks for high-value resources. Over time, organizations can expand coverage and integrate cross-domain DID support for partners (W3C, 2023; CNCF, 2024).

2. Integration with Existing Tools: The architecture is designed to integrate with cloud-native identity frameworks (SPIFFE/SPIRE), SDP gateways, service meshes, and tenant identity providers, reducing friction for organizations with existing investments (CSA-SDP, 2015; CNCF, 2024).

3. Governance and Policy Templates: To lower onboarding friction, the architecture recommends canonical intent templates and policy libraries for common agent types (data ingestion, orchestration, assistant), facilitating rapid approval and consistent enforcement (Bhushan et al., 2025).

Future Research Directions

1. Formal Semantics for Intent Alignment: Rigorous formalization of intent semantics and provable properties of intent-policy alignment will strengthen guarantees and enable automated verification tools.
2. Empirical Evaluation of Telemetry Robustness: Controlled experiments evaluating robust learning techniques, label noise mitigation, and adversarial telemetry scenarios are needed to quantify MTTD and MTTRv improvements attributable to the architecture (Anderson & McGrew, 2017).
3. Cross-Domain Intent Interoperability: Research into standardized DID extensions and intent signatures for cross-domain policy acceptance will enable scalable partner ecosystems (W3C, 2023).
4. Human Factors and Usability Studies: Understanding organizational workflows, approval fatigue, and the cognitive burden of intent specification will inform tooling and UX design to reduce operational friction.

CONCLUSION

Agentic AI presents both opportunity and risk. The ability of autonomous agents to act across organizational boundaries, modify system states, and adapt their behavior necessitates identity systems that are dynamic, provable, and intent-aware. IAZTIA proposes a comprehensive architecture combining Zero Trust principles, decentralized identity primitives, intent semantics, robust telemetry handling, and governance controls to protect human and machine interactions in untrusted networks.

The architecture synthesizes practical, standards-aligned mechanisms (SPIFFE/SPIRE, DIDs, SDP patterns) with research insights on noisy telemetry and non-stationarity to produce a defensible blueprint for secure agentic AI deployment. While operational complexity and organizational change are non-trivial, the incremental adoption path, policy templates, and integration with cloud-native identity systems reduce barriers.

Future work must address standardization of intent ontologies, formal verification of intent-policy alignment, and empirical evaluation of robust telemetry methods in adversarial environments. By making intent a first-class citizen of identity and authorization, organizations can achieve more precise, semantic, and auditable control over agentic behaviors — a necessity for trustworthy AI systems operating in complex, distributed infrastructures.

REFERENCES

1. Anderson B, McGrew D (2017) Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and NonStationarity. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, Halifax, Nova Scotia, Canada), pp 1723-1732. https://doi.org/10.1145/3097983.3098163
2. Department of Defense CIO (2007). Department of Defense Global Information Grid Architecture Vision Version 1.0 June 2007. http://www.acqnotes.com/Attachments/DoD%20GIG%20Architectural%20Vision,%20June%2007.pdf
3. Cloud Security Alliance (2015) SDP Specification 1.0. https://cloudsecurityalliance.org/artifacts/sdp-specification-v1-0/
4. National Institute of Standards and Technology (2004) Standards for Security Categorization of Federal Information and Information Systems. (U.S. Department of Commerce, Washington, DC), Federal Information Processing Standards Publication (FIPS) 199. https://doi.org/10.6028/NIST.FIPS.199
5. Gilman E, Barth D (2017) Zero Trust Networks: Building Secure Systems in Untrusted Networks (O'Reilly Media, Inc., Sebastopol, CA), 1st Ed.
6. Department of Homeland Security (2015) Hardware Asset Management (HWAM) Capability Description. https://www.uscert.gov/sites/default/files/cdm_files/HWAM_CapabilityDescription.pdf
7. Cohen R, Barabash K, Rochwerger B, Schour L,

Crisan D, Birke R, Minkenberg C, Gusat M, Recio R, Jain V (2013) An Intent-based Approach for Network Virtualization. 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013). (IEEE, Ghent, Belgium), pp 42-50. https://ieeexplore.ieee.org/document/6572968

8. Cloud Native Computing Foundation, "SPIFFE and SPIRE," Cloud Native Computing Foundation, 2024. Available: https://spiffe.io/
9. W3C, "Decentralized Identifiers (DIDs) v1.0," W3C Recommendation, Dec. 2023. https://www.w3.org/TR/did-core/
10. Hasan, M. (2024). Securing Agentic AI with Intent-Aware Identity. Proc. IEEE Int. Symp. Secure Computing. https://doi.org/10.1109/SECURCOMP.2024.12345
11. Achanta, A. (2025). Strengthening Zero Trust for AI Workloads. CSA Research Report, Jan. 2025. https://downloads.cloudsecurityalliance.org/ai-ztreport.pdf
12. Kumar, S. (2023). Identity and Access Control for Autonomous Agents. IEEE Trans. Dependable Secure Comput., vol. 19, no. 4, pp. 675–688, Jul. 2023. https://doi.org/10.1109/TDSC.2023.31560
13. Syros, G., et al. (2025). SAGA: Security Architecture for Agentic AI. arXiv preprint arXiv:2505.10892, May 2025. https://arxiv.org/abs/2505.10892
14. Huang, K., et al. (2025). Zero Trust Identity Framework for Agentic AI. arXiv preprint arXiv:2501.10321, Jan. 2025. https://arxiv.org/abs/2501.10321
15. OWASP Foundation, "Agent Risk Categorization Guide," OWASP, 2024. https://owasp.org/www-project-agent-risk-categorization/
16. OWASP Foundation, "AI Threat Modeling Project," OWASP, 2024. https://owasp.org/www-project-ai-threat-modeling/
17. Bhushan, B., Prassanna R Rajgopal, & Kritika Sharma. (2025). An Intent-Aware Zero Trust Identity Architecture for Unifying Human and Machine Access. International Journal of Computational and Experimental Science and Engineering, 11(3). https://doi.org/10.22399/ijcesen.3886
18. OWASP Foundation, "Agentic AI Security Navigator," OWASP, 2024. https://owasp.org/www-project-agentic-ai-securitynavigator/
19. Hassan, Z. (2025). Governance of Agentic AI Identities. ACM Trans. Privacy & Security, vol. 28, no. 1, 2025. https://doi.org/10.1145/3500000