INTERNATIONAL JOURNAL OF NETWORKS AND SECURITY
(ISSN: 2693-387X)
VOL.06 ISSUE 02 (2026)

AMERICAN ACADEMIC PUBLISHER

Research Article

Zero-Trust Identity Federation for Autonomous AI Agents Using SPIFFE and Workload Identity

Ananya Kulkarni[1]

[1]Machine Learning Research Analyst, India

Abstract

Autonomous AI agents increasingly operate across heterogeneous computing environments, including cloud platforms, edge infrastructures, specialized accelerators, and distributed service architectures. This evolution creates an identity-management problem that differs fundamentally from conventional user authentication because autonomous agents may initiate actions, invoke tools, exchange information, and delegate tasks without continuous human intervention. A zero-trust architecture therefore requires identities that are cryptographically verifiable, workload-oriented, short-lived, and independent of network location. This paper proposes a conceptual zero-trust identity federation architecture for autonomous AI agents based on SPIFFE and workload identity principles. The proposed framework separates agent identity, authorization, workload attestation, federation, and policy enforcement while establishing continuous identity verification across agent-to-agent and agent-to-service interactions. The methodology develops an identity lifecycle model, federation mechanism, trust-evaluation process, and policy-enforcement architecture. The analysis further examines how distributed AI workloads and edge environments influence identity requirements. Existing literature demonstrates the growing heterogeneity and distribution of AI computing platforms, but comparatively limited attention has been directed toward identity federation specifically for autonomous agents. The resulting framework provides a structured basis for reducing identity ambiguity, limiting credential persistence, improving cross-domain trust, and supporting policy-driven agent autonomy. The paper concludes that workload-oriented identity can provide an appropriate foundation for zero-trust agent ecosystems, although federation governance, identity revocation, policy consistency, and computational overhead remain significant research challenges.

Keywords: Zero Trust, SPIFFE, Workload Identity, Autonomous AI Agents, Identity Federation, Agentic AI, Distributed AI, Workload Authentication, Trust Management, Cloud-Edge Security.

check for updates

Received: 12 June 2026
Revised: 2 July 2026
Accepted: 20 August 2026
Published: 14 September 2026

Doi:10.55640/ijns-06-02-01
Page No: 40-49

INTRODUCTION

The rapid development of autonomous AI systems is changing the operational model of distributed computing. Modern AI applications increasingly rely on multiple computational components rather than a single centralized model. AI inference can be distributed across cloud systems, edge devices, specialized processors, and application services. Research on edge computing identifies heterogeneity, resource constraints, and distributed execution as important characteristics of contemporary computing environments (Shi et al., 2016; Varghese et al., 2016). Similarly, studies of AI accelerators demonstrate that AI workloads are increasingly executed through heterogeneous hardware architectures designed for specialized computational requirements (Reuther et al., 2021; Mazumder, 2021).

This distributed architecture becomes more challenging when AI components operate as autonomous agents. An autonomous agent can potentially initiate requests, call external services, exchange information with other agents, and perform actions according to dynamically generated objectives. Consequently, traditional identity models based primarily on human users or long-lived application credentials are insufficient for environments in which workloads themselves become active security principals.

Zero-trust security addresses this problem by eliminating implicit trust based on network location or prior connectivity. Every interaction should instead be evaluated according to an authenticated identity, applicable policy, contextual conditions, and the intended operation. For autonomous agents, the critical question is therefore not simply whether an agent belongs to a particular network, but whether the executing workload can prove which identity it represents and whether that identity is authorized to perform a specific operation.

SPIFFE-based zero-trust authentication provides a relevant conceptual foundation for this problem. Pappu, Bhushan, and Mittal (2025) specifically examine SPIFFE-based authentication in AI agent ecosystems, establishing the relevance of workload-oriented identity to autonomous AI environments. The present research extends that perspective by focusing specifically on identity federation, where agents may operate across multiple administrative, cloud, edge, or organizational trust domains.

The primary objective of this paper is to develop a conceptual architecture in which autonomous agents receive verifiable workload identities and use those identities for federated authentication and policy evaluation. The paper focuses on five objectives: establishing a workload-centric identity lifecycle, enabling cross-domain identity federation, supporting continuous zero-trust authentication, integrating authorization with agent identity, and identifying architectural limitations associated with autonomous AI environments.

The significance of the research lies in connecting two rapidly developing architectural trends: autonomous AI workloads and distributed zero-trust infrastructure. As AI systems become more distributed, identity must become portable without becoming uncontrolled. The proposed architecture therefore treats identity as a security control that follows the workload rather than the physical machine or network location.

LITERATURE REVIEW

2.1 Distributed AI and the Identity Problem

The development of AI systems has increasingly involved heterogeneous hardware and distributed execution. Reuther et al. (2021) describe the expanding landscape of AI accelerators, while Mazumder (2021) examines optimization challenges associated with neural-network inference on constrained devices. These studies demonstrate that AI workloads may execute across multiple hardware and software environments rather than a uniform infrastructure.

Edge computing further intensifies this heterogeneity. Shi et al. (2016) characterize edge computing as an architectural paradigm in which computation moves closer to data sources, creating distributed execution conditions. Varghese et al. (2016) similarly identify challenges associated with managing distributed edge resources. From an identity perspective, such environments create a fundamental requirement: security principals must remain identifiable even when workloads migrate between computational locations.

The relevance to autonomous agents is significant. An AI agent may execute reasoning in one environment, call a service in another, and access an edge device or accelerator in a third environment. A machine-oriented identity model can incorrectly associate trust

with infrastructure rather than with the actual workload. Workload identity addresses this conceptual limitation by assigning identity to the executing software workload.

2.2 AI Workload Heterogeneity

AI model architectures also contribute to operational complexity. Vaswani et al. (2017) demonstrated the importance of attention-based architectures in modern neural computation, while Pouyanfar (2018) surveyed deep-learning algorithms and their applications. Model complexity, optimization, and deployment requirements have subsequently encouraged hardware-aware execution.

Deng et al. (2020) examine model compression and hardware acceleration as mechanisms for making neural networks more efficient. Han et al. (2016), Hubara et al. (2016), and Zhu et al. (2017) demonstrate different approaches to reducing model representation and computational requirements. These developments are relevant to identity because optimized AI workloads can be deployed in increasingly diverse execution environments.

A security architecture must therefore avoid assumptions that an AI workload is permanently associated with a specific server, accelerator, or cluster. The identity must remain logically connected to the workload even when its computational substrate changes.

2.3 SPIFFE-Based Authentication for AI Agents

The most directly relevant provided source is the work of Pappu, Bhushan, and Mittal (2025), which explicitly investigates SPIFFE-based zero-trust authentication for AI agent ecosystems. Their work establishes the conceptual relevance of SPIFFE to agent-oriented security and provides the foundation for extending authentication toward federated identity management.

The distinction between authentication and federation is important. Authentication answers whether a workload can establish its identity. Federation additionally determines whether that identity can be recognized across a boundary between different trust domains. Autonomous AI ecosystems require both capabilities because agents may interact with services governed by different security authorities.

2.4 Research Gap

The supplied literature strongly addresses AI computation, hardware acceleration, distributed environments, and workload deployment. However, these studies generally do not develop a comprehensive identity-federation model for autonomous AI agents. The gap is therefore architectural rather than computational.

Existing work supports the need for distributed AI infrastructure, while Pappu et al. (2025) establish the relevance of SPIFFE-based authentication for AI agents. The remaining problem is to integrate workload identity, federated trust, authentication, authorization, and continuous policy evaluation into a unified architecture.

The present paper addresses this gap through a conceptual federation model in which agent identity is independent of network location and can be validated across multiple trust domains.

METHODOLOGY

3.1 Research Design

This study uses a conceptual architecture and analytical research methodology. Rather than conducting a benchmark-based implementation, the research derives architectural

requirements from the characteristics of autonomous AI agents and distributed AI infrastructures. The methodology consists of five stages: identity modeling, workload attestation, federation, authorization, and continuous verification.

The approach is appropriate because the primary research question concerns how identity should be structured for autonomous agents rather than whether a particular implementation achieves a measured performance improvement.

3.2 Workload Identity Model

The proposed architecture assigns each autonomous agent a workload-oriented identity. Conceptually, an agent identity can be represented as:

Agent Identity = Workload + Execution Context + Trust Domain

The workload component identifies the software principal, while the execution context represents the environment in which it is operating. The trust-domain component identifies the authority responsible for issuing or validating the identity.

Unlike a static machine identity, workload identity follows the logical application workload. This is particularly important for AI systems deployed across cloud and edge environments, where computational resources may change dynamically.

For example, an autonomous customer-support agent may operate within Organization A's cloud environment but invoke a translation agent operated by Organization B. The first agent must present an identity that Organization B can verify without requiring direct knowledge of the underlying machine.

3.3 Identity Lifecycle

The proposed lifecycle contains four principal phases: issuance, validation, renewal, and revocation.

During issuance, an identity authority establishes that a workload corresponds to an authorized agent deployment. The identity should have a bounded lifetime so that compromise does not automatically result in permanent access.

During validation, the receiving service verifies the presented workload identity before allowing interaction. Validation must occur independently of network origin.

Renewal allows an actively trusted workload to receive a fresh identity, while revocation terminates trust when the workload, deployment, or policy state becomes unacceptable.

This lifecycle is especially important for autonomous agents because their activity can continue without direct human observation. Pappu et al. (2025) position SPIFFE-based authentication as a mechanism for establishing workload-level trust in AI agent ecosystems; the proposed framework extends this principle by incorporating explicit federation and lifecycle governance.

3.4 Federated Trust Architecture

The architecture contains three logical trust layers.

The first layer is the local workload identity layer, responsible for assigning identities to agents within a trust domain.

The second is the federation layer, which establishes whether identities issued by one domain can be recognized by another.

The third is the policy layer, which determines what an authenticated agent is permitted to do.

Federation should not imply unrestricted trust. Instead, it should establish a controlled mapping between trust domains. For example, Domain A may recognize Domain B's workload identities for read-only data services while rejecting the same identities for administrative operations.

This distinction creates an important separation:

Identity establishes who the workload is; federation establishes whether that identity is recognized; authorization establishes what the workload may do.

3.5 Authentication and Authorization

Authentication is necessary but insufficient for autonomous agents. An authenticated agent could still perform an operation beyond its intended authority.

The proposed framework therefore associates workload identities with policy attributes. These may include agent role, service type, requested operation, trust domain, and execution state.

Consider an AI planning agent that is authorized to retrieve inventory information but not modify inventory records. Even if the agent possesses a valid workload identity, the authorization layer should reject a write operation.

This separation follows zero-trust reasoning and prevents identity possession from becoming equivalent to unrestricted access.

3.6 Continuous Verification

Autonomous-agent interactions should be evaluated continuously rather than assuming that an earlier successful authentication remains permanently valid. A request can therefore be evaluated using the following conceptual decision model:

Access Decision = Identity Validity × Federation Trust × Policy Authorization × Contextual Risk

If any critical component fails, the request should be rejected or subjected to additional controls.

This model is particularly relevant to distributed AI because workload conditions may change after initial authentication. An agent could migrate, change software state, or interact with a new service. Continuous verification therefore reduces dependence on static trust assumptions.

3.7 Agent-to-Agent Federation

A distinctive feature of the proposed architecture is direct agent-to-agent authentication.

Suppose Agent A requires a specialized reasoning service provided by Agent B. Agent A presents its workload identity. Agent B validates the identity against an accepted trust domain and evaluates whether Agent A's requested operation is authorized.

If Agent B subsequently calls Agent C, B does not automatically transfer A's authority to C. Instead, each interaction generates an independent authentication and authorization decision. This prevents uncontrolled transitive trust.

Such compartmentalization is essential for autonomous systems because an agent may

dynamically construct multi-agent workflows.

3.8 Integration with Edge and Accelerated AI

Distributed AI systems frequently use specialized hardware and edge environments. Reuther et al. (2021), Lim et al. (2021), Wu et al. (2020), and Liu et al. (2020) demonstrate the importance of hardware-aware AI execution. Ray (2022) similarly discusses TinyML and constrained AI inference.

The proposed identity architecture treats hardware as an execution substrate rather than the ultimate security principal. Consequently, an AI workload running on an accelerator should retain its logical identity even if its computational environment changes.

This design reduces the risk of coupling authorization policies to infrastructure topology. It also provides a more consistent security model for cloud-edge AI deployments.

3.9 Threat Model

The conceptual threat model considers compromised agents, stolen credentials, unauthorized workloads, malicious cross-domain requests, identity impersonation, and excessive delegation.

The architecture mitigates these risks through short-lived workload identities, independent federation validation, least-privilege authorization, and continuous verification.

However, identity alone cannot prevent malicious behavior by a legitimate agent. An authenticated agent could still make unsafe decisions. Consequently, workload identity should be treated as a foundational control rather than a complete AI-security solution.

RESULTS

The analytical evaluation produces four primary findings.

First, workload identity is more suitable than infrastructure identity for autonomous AI ecosystems. Distributed AI environments increasingly involve heterogeneous processors, edge systems, and specialized accelerators. Research on AI acceleration and edge computing demonstrates that computational workloads are not necessarily tied to one stable infrastructure location (Reuther et al., 2021; Shi et al., 2016). A workload-oriented identity consequently provides stronger architectural continuity than machine-centric identity.

Second, federation must remain distinct from authentication. Authentication establishes an agent's identity, but federated environments require an additional trust decision concerning whether an external identity authority is recognized. The SPIFFE-oriented approach discussed by Pappu et al. (2025) provides a foundation for workload authentication, while the proposed extension adds explicit cross-domain trust governance.

Third, identity federation should not create automatic transitive authorization. If Agent A is trusted by Agent B and Agent B is trusted by Agent C, it does not logically follow that A should receive B's privileges within C. Each interaction should therefore produce an independent authorization decision. This separation reduces privilege propagation across autonomous workflows.

Fourth, distributed AI deployment increases the importance of portable identity. AI systems can be optimized for different computational environments, as reflected in research on model compression, hardware acceleration, TinyML, and specialized neural-network processors (Han et al., 2016; Deng et al., 2020; Ray, 2022). Portable workload

identity allows security policy to remain stable while execution infrastructure changes.

The findings also reveal an important limitation: identity federation cannot independently determine whether an agent's behavior is safe. A valid identity confirms a security principal but does not prove that its generated decision is correct or desirable. Consequently, future architectures must combine identity with behavioral monitoring, contextual authorization, and stronger agent-governance mechanisms.

Overall, the analysis indicates that SPIFFE-aligned workload identity provides a technically appropriate foundation for federated zero-trust authentication, but the architecture must extend beyond identity issuance toward lifecycle management, policy enforcement, and cross-domain governance. This conclusion is consistent with the central security problem identified for AI agent ecosystems by Pappu et al. (2025).

DISCUSSION

The principal theoretical contribution of this research is the separation of identity, federation, and authorization within autonomous AI security. Conventional security designs can implicitly combine these concepts by treating possession of a credential as sufficient evidence of authority. Autonomous agents require a stricter model because they can independently initiate interactions and construct multi-step workflows.

The proposed architecture establishes identity as the first security boundary. Federation then becomes a controlled mechanism for recognizing external identities rather than an unrestricted trust relationship. Authorization subsequently determines the precise operation allowed. This layered approach supports the zero-trust principle that trust should be evaluated at each interaction.

The architecture also addresses a fundamental characteristic of distributed AI: infrastructure heterogeneity. AI accelerators, edge systems, and constrained devices have different computational characteristics (Mazumder, 2021; Varghese et al., 2016). Security identity should therefore remain independent of the computational substrate. If identity were directly attached to hardware, workload migration could require policy reconstruction and increase operational complexity.

A further implication concerns autonomous delegation. Agentic systems may delegate tasks to other agents. Without explicit identity boundaries, delegated authority can propagate unexpectedly. The proposed independent authorization model prevents a receiving agent from assuming that the privileges of the initiating agent automatically apply to subsequent interactions.

The framework nevertheless introduces trade-offs. Short-lived identity credentials improve security but increase renewal operations. Federation improves interoperability but introduces governance dependencies between trust domains. Continuous verification improves control but may introduce latency and computational overhead, particularly on constrained edge systems. Such trade-offs are important given the resource limitations documented in research on TinyML and edge computing (Ray, 2022; Shi et al., 2016).

Another limitation concerns compromised legitimate workloads. SPIFFE-based authentication can establish that a workload corresponds to an authorized identity, but it cannot guarantee that the workload's internal decision-making is benign. Pappu et al. (2025) provide the immediate conceptual basis for applying SPIFFE authentication to AI agent ecosystems, but the broader problem of behavioral trust remains open.

The literature also reveals a disciplinary gap. The provided AI literature concentrates predominantly on computational efficiency, accelerator architecture, model compression, and edge deployment. These studies are highly relevant to understanding where AI workloads execute, but they provide limited treatment of federated workload

identity. The proposed research therefore connects infrastructure architecture with security architecture rather than treating them as independent concerns.

From a practical perspective, organizations deploying multi-agent AI systems should avoid treating network placement as a substitute for identity. An agent operating inside a trusted cloud network should not automatically receive more trust than an externally hosted agent if both can present equivalent, verifiable workload identities. Conversely, federation should permit carefully controlled external collaboration without requiring every organization to share its internal identity infrastructure.

Future implementations should evaluate the framework through measurable dimensions including authentication latency, identity-renewal overhead, federation scalability, policy-evaluation cost, revocation propagation time, and resistance to credential compromise. Such empirical evaluation would complement the conceptual results presented here.

CONCLUSION

This paper proposed a conceptual zero-trust identity federation architecture for autonomous AI agents using SPIFFE and workload identity principles. The analysis demonstrates that autonomous agents require a security model centered on workloads rather than users, machines, or network locations. This requirement becomes increasingly important as AI workloads move across heterogeneous cloud, edge, and accelerated computing environments.

The proposed framework separates four fundamental security functions: workload identity issuance, cross-domain federation, authorization, and continuous verification. This separation prevents authentication from being interpreted as unrestricted authority and reduces the possibility of privilege propagation across autonomous agent workflows.

The research also identifies an important distinction between computational scalability and security scalability. AI infrastructures may successfully scale models and hardware while still lacking an identity architecture capable of managing thousands of autonomous workloads. Federation therefore becomes an essential architectural layer for multi-domain agent ecosystems.

The principal contribution is a structured model in which agent identity remains portable across execution environments while trust remains explicitly governed by policy. The approach builds upon the SPIFFE-oriented AI-agent authentication perspective presented by Pappu et al. (2025) and extends it toward federated cross-domain operation.

Several limitations remain. The framework is conceptual rather than experimentally validated, and the supplied literature provides limited empirical evidence regarding SPIFFE federation performance in large-scale agentic AI systems. Furthermore, workload identity cannot by itself establish behavioral correctness or prevent a legitimate but compromised agent from making harmful decisions.

Future research should therefore investigate experimental implementations of federated SPIFFE infrastructures for multi-agent AI systems, focusing on latency, scalability, revocation, policy consistency, delegation control, and resilience against compromised workloads. Integrating workload identity with behavioral trust and adaptive authorization represents another important research direction.

Ultimately, zero-trust identity federation can provide a foundational security layer for autonomous AI ecosystems by ensuring that every agent interaction is attributable, verifiable, and policy-controlled. As AI systems become increasingly autonomous and distributed, portable workload identity is likely to become a central architectural

requirement for secure agent-to-agent and agent-to-service communication.

REFERENCES

1. Karine, T. Napoléon, J.-Y. Mulot, and Y. Auffret, "Video seals recognition using transfer learning of convolutional neural network," in Proc. 10th Int. Conf. Image Process. Theory, Tools, Appl., 2020, pp. 1–4.

2. A. N. Mazumder, "A survey on the optimization of neural network accelerators for micro-AI on-device inference," IEEE Trans. Emerg. Sel. Topics Circuits Syst., vol. 11, no. 4, pp. 532–547, Dec. 2021.

3. A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "AI accelerator survey and trends," in Proc. IEEE High Perform. Extreme Comput. Conf., 2021, pp. 1–9.

4. A. Vaswani, "Attention is all you need," in Proc. Int. Conf. Neural Inf. Process. Syst., 2017, vol. 30, pp. 6000–6010.

5. B. Varghese, N. Wang, S. Barbhuiya, P. Kilpatrick, and D. S. Nikolopoulos, "Challenges and opportunities in edge computing," in Proc. IEEE Int. Conf. Smart Cloud, 2016, pp. 20–26.

6. C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," in Proc. 5th Int. Conf. Learn. Representations, Toulon, France, 2017.

7. E. Manor and S. Greenberg, "Custom hardware inference accelerator for tensorflow lite for microcontrollers," IEEE Access, vol. 10, pp. 73484–73493, 2022.

8. E. Wang, "Deep neural network approximation for custom hardware: Where we've been, where we're going," ACM Comput. Surv., vol. 52, no. 2, pp. 1–39, 2019.

9. I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," Proc. Adv. Neural Inform. Process. Syst., D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds., vol. 29, 2016.

10. K. Pappu, B. Bhushan and A. Mittal, "SPIFFE-Based Zero-Trust Authentication for AI Agent Ecosystems," 2025 International Conference on Computer and Applications (ICCA), Bahrain, Bahrain, 2025, pp. 1-7, doi: 10.1109/ICCA66035.2025.11431026.

11. L. Deng, G. Li, S. Han, L. Shi, and Y. Xie, "Model compression and hardware acceleration for neural networks: A comprehensive survey," Proc. IEEE, vol. 108, no. 4, pp. 485–532, Apr. 2020.

12. L. Sekanina, "Neural architecture search and hardware accelerator co-search: A survey," IEEE Access, vol. 9, pp. 151337–151362, 2021.

13. N. Wu, T. Jiang, L. Zhang, F. Zhou, and F. Ge, "A reconfigurable convolutional neural network-accelerated coprocessor based on RISC-V instruction set," Electronics, vol. 9, no. 6, 2020, Art. no. 1005.

14. P. P. Ray, "A review on TinyML: State-of-the-art and prospects," J. King Saud Univ.-Comput. Inf. Sci., vol. 34, no. 4, pp. 1595–1623, 2022.

15. S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in Proc. 4th Int. Conf. Learn. Representations, Y. Bengio and Y. LeCun, eds., San Juan, Puerto Rico, 2016.

16. S. Mittal, "A survey of FPGA-based accelerators for convolutional neural networks," Neural Comput. Appl., vol. 32, no. 4, pp. 1109–1139, 2020.

17. S. Pouyanfar, "A survey on deep learning: Algorithms, techniques, and applications," ACM Comput. Surv., vol. 51, no. 5, pp. 1–36, 2018.

18. S.-H. Lim, W. W. Suh, J.-Y. Kim, and S.-Y. Cho, "RISC-V virtual platform-based convolutional neural network accelerator implemented in systemC," Electronics, vol. 10, no. 13, 2021, Art. no. 1514.

19. V. Murahari, C. E. Jimenez, R. Yang, and K. Narasimhan, "DataMUX: Data multiplexing for neural networks," in Proc. Adv. Neural Inform. Process. Syst., A. H. Oh, A. Agarwal, D. Belgrave, and Kyunghyun Cho, eds., 2022.

20. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016.

21. Z. Liu, J. Jiang, G. Lei, K. Chen, B. Qin, and X. Zhao, "A heterogeneous processor design for CNN-based AI applications on IoT devices," Procedia Comput. Sci., vol. 174, pp. 2–8, 2020.