COLOMBO SCIENTIFIC PUBLISHING

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE & INFORMATION SYSTEM
E-ISSN: 2536-7919
P-ISSN: 2536-7900
https://doi.org/10.55640/ijcsis/Volume11Issue09-04 PAGE NO: 27-34

Identity-Aware Authentication for Agentic AI Systems: A Taxonomy of Protocols, Threat Models, and Trust Mechanisms

Faisal Al-Harbi
Department Artificial
Intelligence Research Specialist, Saudi Arabia

ARTICLE INFO

Article history:

Submission: July 01, 2026
Accepted: August 17, 2026
Published: September 14, 2026
VOLUME: Vol.11 Issue 09 2026

Keywords:
Agentic AI, Identity-Aware Authentication, AI Agents, Authentication Protocols, Trust Management, Threat Models, Multi-Agent Systems, Behavioral Authentication, Reinforcement Learning, Neuro-Symbolic AI

ABSTRACT

The emergence of agentic artificial intelligence (AI) systems introduces a significant transformation in the conventional understanding of digital identity and authentication. Unlike conventional software applications, agentic AI systems can perceive environments, reason over contextual information, invoke external tools, communicate with other agents, and execute actions with varying degrees of autonomy. Consequently, authentication must establish not only the identity of an initiating entity but also the identity, authority, behavioral context, and trustworthiness of an autonomous agent acting on behalf of a user or organization. This paper develops an identity-aware authentication taxonomy for agentic AI ecosystems by synthesizing concepts from authentication, reinforcement learning, multi-agent systems, safe AI, causal reasoning, and neuro-symbolic computing. The methodology organizes authentication mechanisms according to identity representation, protocol interaction, contextual verification, threat model, and trust decision. The proposed framework distinguishes static identity authentication, delegated identity authentication, contextual authentication, continuous behavioral authentication, and multi-agent trust authentication. It further categorizes threats into impersonation, credential misuse, agent substitution, delegation abuse, behavioral manipulation, and inter-agent trust exploitation. The analysis demonstrates that authentication for agentic AI cannot be adequately represented as a one-time binary verification event. Instead, authentication should operate as a continuous, risk-aware decision process that integrates identity evidence, contextual signals, authorization constraints, behavioral observations, and trust evolution. The paper contributes a conceptual taxonomy and research framework for designing identity-aware authentication mechanisms capable of supporting autonomous and multi-agent AI environments.

INTRODUCTION

The development of autonomous and semi-autonomous AI systems is changing the security assumptions traditionally associated with software applications. Conventional authentication generally establishes whether an identified subject is permitted to access a resource. Agentic AI introduces an additional problem: the authenticated subject may itself be an autonomous computational entity capable of making decisions, invoking services, communicating with other agents, and executing actions over time. Consequently, authentication must establish not only who is requesting access but also which agent instance is acting, on whose behalf, under what delegation, and within which operational context.

This distinction becomes particularly important in environments involving multiple autonomous agents. Multi-agent reinforcement learning demonstrates that computational agents can interact within dynamic environments and influence one another through coordinated decision-making (Buşoniu, Babuška, and De Schutter, 2010). Similarly, reinforcement learning research establishes that an intelligent system can adapt its behavior according to environmental feedback rather than following only predetermined execution sequences (Kaelbling, Littman, and Moore, 1996). These characteristics create a security challenge because the behavior of an authenticated agent may evolve after authentication.

The conceptual problem of authentication in agentic AI ecosystems has been explicitly examined by Bhushan, Pappu, and Mittal, who position authentication as an architectural and protocol-level concern within agentic AI ecosystems and emphasize systematic classification of authentication approaches (Bhushan, Pappu, and Mittal, 2025). Their framework provides an important foundation for examining identity-aware authentication as a distinct security problem rather than simply applying conventional authentication mechanisms to AI agents.

The central problem addressed by this paper is therefore the absence of a unified taxonomy that connects agent identity, authentication protocols, threat models, behavioral evidence, and trust mechanisms. Existing work on reinforcement learning, safe reinforcement learning, multi-agent learning, causal reinforcement learning, and neuro-symbolic reasoning provides relevant theoretical components, but these components have generally been studied as AI capabilities rather than as an integrated authentication architecture.

The objective of this research is to develop a conceptual identity-aware authentication taxonomy for agentic AI systems. The study specifically examines how authentication can incorporate persistent identity, delegated identity, contextual evidence, behavioral observations, multi-agent relationships, and dynamically changing trust. The scope is conceptual rather than implementation-specific. The resulting framework is intended to provide a foundation for protocol design, threat analysis, and future empirical evaluation of authentication mechanisms for autonomous AI ecosystems.

LITERATURE REVIEW

Research on intelligent agents provides an important theoretical foundation for understanding why authentication requirements change when software entities become autonomous. Kaelbling, Littman, and Moore (1996) characterize reinforcement learning as a framework in which an agent learns through interaction with its environment. This interaction-oriented model implies that an agent cannot always be understood exclusively through its initial configuration because subsequent behavior depends on environmental conditions and accumulated experience.

Arulkumaran et al. (2017) further explain the development of deep reinforcement learning, demonstrating how learning systems can combine representation learning with decision-making. Li (2017) similarly provides an overview of deep reinforcement learning and its ability to support complex decision processes. These studies are not authentication studies, but they establish a significant security implication: an AI agent may produce behavior that is dynamically generated rather than completely predetermined. Therefore, authentication based solely on an initial identity assertion may become insufficient for long-lived autonomous agents.

The multi-agent perspective adds another dimension. Buşoniu, Babuška, and De Schutter (2010) examine multi-agent reinforcement learning as an environment in which several agents interact and learn. In such systems, identity becomes relational. An agent may need to establish not only its own identity but also the identity of the agent with which it is communicating and the legitimacy of the relationship between them. This creates a requirement for inter-agent authentication and trust establishment.

Safety-oriented reinforcement learning provides another theoretical connection. García and Fernández (2015) examine safe reinforcement learning and highlight the importance of controlling learning behavior under safety constraints. For identity-aware authentication, the implication is that an authenticated agent should not automatically be considered trustworthy for every subsequent action. Authentication establishes identity evidence, whereas trust mechanisms must evaluate whether the observed behavior remains consistent with permitted operational boundaries.

Padakandla (2021) emphasizes reinforcement learning in dynamically varying environments. Such dynamic environments are directly relevant to continuous authentication because contextual conditions may change after an initial authentication event. A system may therefore need to reassess identity confidence when the agent changes location, tools, interaction patterns, objectives, or communication partners.

The neuro-symbolic literature contributes an additional theoretical perspective. Besold (2017) examines neural-symbolic learning and reasoning, while Wang and Yang (2022) and Yu et al. (2021) discuss the integration of data-driven and symbolic approaches. These approaches are relevant because identity-aware authentication requires both statistical evidence and explicit security rules. A behavioral model can detect anomalous activity, whereas symbolic policies can express constraints such as permitted delegation, resource restrictions, and agent-role relationships.

Causal reinforcement learning provides another potentially important mechanism. Zeng et al. (2023) examine causal relationships in reinforcement learning, suggesting a conceptual basis for distinguishing correlation from causally meaningful behavioral evidence. Within authentication, this distinction can help determine whether an unusual agent action is genuinely associated with identity compromise or merely reflects an environmental change.

Silver et al. (2016) and Liu et al. (2021) demonstrate the increasing sophistication of AI systems capable of planning, searching, and decision-making in complex environments. As agent capabilities increase, the consequences of identity compromise also increase. An attacker that impersonates a simple application may obtain data, whereas impersonating an autonomous agent may enable the attacker to initiate a sequence of actions.

Uprety and Rawat (2021) specifically connect reinforcement learning with IoT security, illustrating the relevance of adaptive learning approaches to security decisions. This supports the argument that authentication for autonomous environments can benefit from adaptive security mechanisms but also demonstrates the importance of controlling learning behavior within security-sensitive contexts.

The most directly relevant contribution is the work of Bhushan, Pappu, and Mittal (2025), which provides a conceptual framework for authentication in agentic AI ecosystems and emphasizes protocol analysis and taxonomy. Building on this foundation, the present study extends the discussion toward identity-aware authentication by connecting identity representation with contextual verification, behavioral evidence, threat models, and evolving trust.

The literature therefore reveals a research gap. Existing studies provide substantial foundations for autonomous decision-making, multi-agent interaction, safe learning, neuro-symbolic reasoning, and adaptive security, but these concepts are not sufficiently integrated into a single authentication model. This paper addresses that gap through a structured taxonomy.

METHODOLOGY

3.1 Research Design

This study adopts a conceptual research methodology based exclusively on structured synthesis of the supplied literature. The methodology does not introduce external datasets or empirical measurements. Instead, the provided studies are mapped according to their relevance to five authentication dimensions: identity, protocol, context, threat, and trust.

The approach is motivated by the observation that authentication in agentic AI is multidimensional. A traditional authentication process can often be represented as a function that evaluates credentials and produces an allow-or-deny outcome. For an autonomous agent, however, the authentication decision can be represented conceptually as:

$$A = f(I, P, C, B, D, T, R)$$

where I represents identity evidence, P represents protocol evidence, C represents contextual information, B represents behavioral evidence, D represents delegation information, T represents accumulated trust, and R represents current risk.

This representation does not prescribe a particular algorithm. Instead, it establishes the analytical dimensions required to understand identity-aware authentication.

3.2 Taxonomy of Identity Models

The first category is static agent identity. Under this model, an agent possesses a persistent identifier associated with cryptographic or system-level credentials. The primary objective is to determine whether the presented identity corresponds to a recognized agent.

The second category is delegated identity. Here, the agent acts on behalf of another principal, such as a user, service, or organizational entity. Authentication must consequently establish two relationships: the identity of the agent and the legitimacy of the delegation relationship. Bhushan, Pappu, and Mittal (2025) provide an important conceptual basis for treating authentication in agentic ecosystems as a broader protocol problem rather than a simple credential check.

The third category is context-aware identity. Identity confidence is evaluated together with contextual information. An authenticated agent operating within its expected environment may receive a stronger confidence assessment than the same identity exhibiting unexpected contextual characteristics.

The fourth category is behavioral identity. Here, the system evaluates whether the agent's observed actions remain consistent with its established behavioral profile. Reinforcement learning research demonstrates that agent behavior can evolve through interaction (Kaelbling, Littman, and Moore, 1996). Therefore, behavioral authentication should be considered complementary to initial identity verification rather than a replacement for it.

The fifth category is relational identity, in which identity is evaluated in the context of interactions among multiple agents. This is particularly important for multi-agent systems, where an agent must establish the identity and trust status of other participating agents (Buşoniu, Babuška, and De Schutter, 2010).

3.3 Authentication Protocol Taxonomy

The proposed taxonomy identifies four functional protocol stages.

The first is initial identity establishment, where an agent presents evidence of its identity. The second is delegation verification, where the system determines whether the agent is authorized to act on behalf of another entity. The third is contextual verification, where environmental and operational characteristics are examined. The fourth is continuous verification, where identity confidence is updated during agent operation.

A hypothetical example illustrates the distinction. Suppose an AI agent is authenticated to access a corporate data repository. Static authentication may confirm the agent's identity. Delegation verification may confirm that it is acting for a specific user. Contextual verification may determine whether the requested operation is consistent with the assigned task. Continuous verification may identify whether subsequent behavior deviates significantly from the agent's expected operational profile.

This layered design is consistent with the conceptual direction of authentication taxonomies for agentic ecosystems proposed by Bhushan, Pappu, and Mittal (2025), while extending the analysis toward continuous identity confidence.

3.4 Threat Model

The threat taxonomy consists of six major categories.

Impersonation attacks occur when an adversary attempts to represent itself as a legitimate AI agent.

Credential compromise occurs when legitimate authentication material is acquired by an unauthorized entity.

Agent substitution occurs when an authenticated agent is replaced or redirected by another computational entity.

Delegation abuse occurs when legitimate delegation is extended beyond its intended scope.

Behavioral manipulation occurs when an attacker causes an authenticated agent to produce actions inconsistent with its legitimate operating constraints.

Inter-agent trust exploitation occurs when one malicious or compromised agent manipulates the trust relationship established with another agent.

The distinction between identity compromise and behavioral manipulation is particularly important. A compromised identity may still pass a conventional authentication check, while its subsequent behavior reveals the security violation. Safe reinforcement learning provides a theoretical foundation for considering operational constraints alongside adaptive behavior (García and Fernández, 2015).

3.5 Trust Mechanism Model

Authentication and trust are treated as related but distinct concepts. Authentication answers the question: "Is this entity sufficiently verified as the claimed identity?" Trust answers the broader question: "Given available evidence, how confidently should this entity be allowed to perform a particular action?"

A conceptual trust score can therefore be expressed as:

$$T(t+1) = g(T(t), E_i, B_i, C_i, V_i, R_i)$$

where the trust state at time t+1 depends on previous trust, identity evidence, behavioral observations, contextual evidence, verification outcomes, and risk.

Neuro-symbolic approaches provide a useful theoretical basis for combining learned evidence with explicit rules. Neural models can identify behavioral deviations, while symbolic rules can enforce non-negotiable security constraints (Besold, 2017; Wang and Yang, 2022). This hybrid model is particularly appropriate where autonomous agents operate in environments requiring both flexibility and policy compliance.

3.6 Continuous and Adaptive Authentication

Continuous authentication is essential when agents operate for extended periods. Padakandla (2021) highlights the importance of adaptive learning under dynamically varying conditions. Authentication mechanisms should therefore distinguish legitimate behavioral adaptation from suspicious behavioral deviation.

Causal reasoning may further improve this process. Zeng et al. (2023) provide a foundation for reasoning about causal relationships within reinforcement learning. In authentication, causal analysis could help distinguish an agent's abnormal behavior caused by legitimate environmental changes from behavior caused by identity compromise.

The methodology consequently treats authentication as a dynamic process rather than a single event. This model is particularly appropriate for highly capable systems demonstrated in advanced reinforcement-learning applications (Silver et al., 2016; Liu et al., 2021).

RESULTS

The conceptual synthesis produces five principal findings.

First, agent identity must be separated from agent authorization and trust. An authenticated identity does not automatically imply that every action performed by the agent is trustworthy. This distinction becomes particularly important in autonomous systems because agent behavior can change as environmental

conditions and learned policies evolve. Reinforcement learning literature establishes the adaptive nature of agent decision-making (Kaelbling, Littman, and Moore, 1996; Arulkumaran et al., 2017).

Second, authentication should become contextual and continuous. Static identity verification is valuable for establishing an initial security boundary, but it provides limited protection against subsequent compromise or behavioral manipulation. Dynamic environments discussed in reinforcement-learning research support the need for authentication mechanisms capable of updating confidence over time (Padakandla, 2021).

Third, multi-agent environments require relational authentication. In an autonomous ecosystem, an agent does not operate in isolation. It may receive information from, delegate tasks to, or request services from other agents. Consequently, the identity of both participants and the legitimacy of their relationship must be evaluated. Multi-agent reinforcement learning provides the theoretical basis for this relational perspective (Buşoniu, Babuška, and De Schutter, 2010).

Fourth, hybrid neural-symbolic security mechanisms are promising. Learned models can identify complex behavioral patterns, while symbolic mechanisms can encode explicit identity, delegation, and authorization constraints. The neuro-symbolic literature supports the conceptual feasibility of combining data-driven inference with structured reasoning (Besold, 2017; Wang and Yang, 2022; Yu et al., 2021).

Fifth, trust should be treated as an evolving security state rather than a fixed attribute. A legitimate agent may accumulate trust through consistent behavior, while anomalous behavior may reduce confidence. Safe reinforcement learning reinforces the importance of constraining adaptive behavior within safety boundaries (García and Fernández, 2015).

The findings therefore support a layered identity-aware architecture consisting of identity establishment, delegation validation, contextual assessment, continuous behavioral verification, and dynamic trust evaluation. This architecture is consistent with the broader authentication taxonomy proposed for agentic AI ecosystems by Bhushan, Pappu, and Mittal (2025), while emphasizing the additional importance of continuous and behavioral identity assessment.

The resulting taxonomy also indicates that authentication failures can originate at different levels. An identity credential can be valid while delegation is illegitimate; delegation can be legitimate while behavior becomes anomalous; and behavior can appear unusual without necessarily indicating compromise. Therefore, effective authentication requires correlation among multiple evidence types rather than dependence on a single security signal.

DISCUSSION

The principal theoretical implication of this study is that authentication for agentic AI should be conceptualized as a stateful security process rather than a discrete credential-validation event. Conventional models often assume that successful authentication establishes a sufficiently stable security relationship. Agentic systems challenge this assumption because autonomy introduces behavioral variability, learning, adaptation, and inter-agent interaction.

The distinction between authentication and trust is therefore fundamental. Authentication provides evidence concerning identity, while trust incorporates identity evidence with behavioral, contextual, and historical information. Treating the two concepts as identical can produce excessive confidence in authenticated agents. This concern is particularly relevant to adaptive systems whose policies may change through interaction, as demonstrated by reinforcement-learning research (Arulkumaran et al., 2017; Padakandla, 2021).

The framework also exposes an important trade-off between security and autonomy. Strong continuous verification can reduce the risk associated with compromised agents, but excessive verification may interrupt legitimate autonomous operation. An agent executing a long sequence of legitimate actions should not necessarily be forced through an equivalent authentication procedure before every operation. A risk-

adaptive architecture is therefore preferable, in which verification intensity increases when identity confidence decreases or operational risk increases.

Multi-agent environments introduce another contradiction. Trust is necessary for efficient collaboration, but excessive trust creates opportunities for cascading compromise. An agent may authenticate another agent correctly while being unable to determine whether that agent has itself been compromised. Multi-agent reinforcement learning demonstrates the complexity of interaction among autonomous entities (Buşoniu, Babuška, and De Schutter, 2010). This suggests that inter-agent trust should be bounded, revocable, and context-specific.

The proposed neural-symbolic direction also involves a trade-off. Learned behavioral models can recognize complex patterns that are difficult to specify manually, while symbolic policies provide explicit and auditable security constraints. However, learned models may introduce uncertainty, whereas symbolic policies can become rigid in rapidly changing environments. Combining both mechanisms therefore requires careful conflict-resolution policies (Besold, 2017; Yu et al., 2021).

Causal reasoning provides a further opportunity but also a limitation. If authentication systems can distinguish the causes of behavioral changes, they may reduce false positives. Nevertheless, causal inference in complex autonomous environments remains difficult because numerous environmental variables can influence agent behavior. The causal reinforcement-learning literature indicates the relevance of this direction but does not by itself provide a complete authentication solution (Zeng et al., 2023).

From a practical perspective, the framework suggests that future agentic AI platforms should maintain explicit relationships among identity, delegation, authorization, behavioral history, and trust. The conceptual framework developed by Bhushan, Pappu, and Mittal (2025) is particularly relevant because it establishes authentication as a specialized concern for agentic AI ecosystems rather than merely an extension of conventional application security.

The principal limitation of this research is its conceptual nature. The framework has not been validated using operational datasets, attack simulations, or benchmark authentication protocols. Furthermore, the supplied literature is stronger in reinforcement learning, multi-agent intelligence, and neuro-symbolic reasoning than in empirical authentication evaluation. Consequently, the proposed taxonomy should be interpreted as a research architecture rather than a demonstrated production system.

Despite these limitations, the synthesis provides a coherent theoretical foundation for future experimental research. In particular, future studies can evaluate how identity confidence should be updated, how trust decay should be modeled, how behavioral anomalies should influence authentication decisions, and how multi-agent trust can be contained when one participating agent is compromised.

CONCLUSION

This paper developed an identity-aware authentication taxonomy for agentic AI systems by integrating concepts from authentication, reinforcement learning, multi-agent systems, safe learning, causal reasoning, and neuro-symbolic AI. The analysis demonstrates that autonomous agents create authentication requirements that extend beyond conventional identity verification.

The central contribution is the conceptual separation of five related dimensions: identity establishment, authentication protocols, contextual verification, threat modeling, and dynamic trust. Static authentication remains necessary but is insufficient for long-lived autonomous agents. Delegated identity requires verification of both the agent and the authority under which it operates. Multi-agent environments require relational authentication, while adaptive behavior motivates continuous verification.

The literature further suggests that hybrid neural-symbolic mechanisms can provide an effective conceptual balance between behavioral inference and explicit security policy. Reinforcement-learning research supports the need to account for adaptive behavior, whereas safe reinforcement learning emphasizes operational constraints. Multi-agent research demonstrates the importance of inter-agent

relationships, and causal reinforcement learning provides a potential basis for distinguishing legitimate behavioral changes from security-relevant anomalies.

The study builds directly upon the authentication framework proposed by Bhushan, Pappu, and Mittal (2025) and extends its conceptual direction toward continuous identity confidence and trust-aware authentication. The resulting framework treats authentication as an evolving security state rather than a single event.

REFERENCES

1. K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” IEEE Signal Process. Mag., vol. 34, no. 6, pp. 26–38, Nov. 2017.

2. T. R. Besold, “Neural-symbolic learning and reasoning: A survey and interpretation,” 2017, arXiv:1711.03902.

3. D. Bouneffouf and C. C. Aggarwal, “Survey on applications of neurosymbolic artificial intelligence,” 2022, arXiv:2209.12618.

4. L. Buşoniu, R. Babuška, and B. De Schutter, “Multi-agent reinforcement learning: An overview,” in Innovations in Multi-Agent Systems and Applications-1. Berlin, Germany : Springer, 2010, pp. 183–221.

5. B. Bhushan, K. Pappu and A. Mittal, "A Conceptual Framework for Authentication in Agentic AI Ecosystems: Protocol Analysis and Taxonomy," 2025 International Conference on Computer and Applications (ICCA), Bahrain, Bahrain, 2025, pp. 1-8, doi: 10.1109/ICCA66035.2025.11430864.

6. J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” J. Mach. Learn. Res., vol. 16, no. 1, pp. 1437–1480, 2015.

7. L. P. Kaelbling, M. L. Littman, and A. W. Moore, “Reinforcement learning: A survey,” J. Artif. Intell. Res., vol. 4, pp. 237–285, 1996.

8. Y. Li, “Deep reinforcement learning: An overview,” 2017, arXiv:1701.07274.

9. R.-Z. Liu, W. Wang, Y. Shen, Z. Li, Y. Yu, and T. Lu, “An introduction of mini-alphastar,” 2021, arXiv:2104.06890.

10. S. Padakandla, “A survey of reinforcement learning algorithms for dynamically varying environments,” ACM Comput. Surv., vol. 54, no. 6, pp. 1–25, 2021.

11. D. Silver, “Mastering the game of go with deep neural networks and tree search,” Nature, vol. 529, no. 7587, pp. 484–489, 2016.

12. A. Uprety and D. B. Rawat, “Reinforcement learning for IoT security: A comprehensive survey,” IEEE Internet Things J., vol. 8, no. 11, pp. 8693–8706, Jun. 2021.

13. W. Wang and Y. Yang, “Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing,” 2022, arXiv:2210.15889.

14. D. Yu, B. Yang, D. Liu, and H. Wang, “A survey on neural-symbolic systems,” 2021, arXiv:2111.08164.

15. Y. Zeng, R. Cai, F. Sun, L. Huang, and Z. Hao, “A survey on causal reinforcement learning,” 2023, arXiv:2302.05209.