Indian Engineering Journal
ISSN: 2984-9047
Vol 4 Issue 2 (2026)

Secure Multi-Agent Systems: Authentication, Trust, and Privacy

Dr. T. Sujithra, Dr. A. Indrapandi, R. Abhishek

[1]Assistant Professor, Department of Computer Science, Mannar Thirumalai Naicker College Madurai

sujithra.thangam@gmail.c om

[2]Assistant Professor, Department of Computer Application, The American College, Madurai, Tamilnadan, India

indrapandiaaa@gmail.com

[3]Student, Department of Computer Application, The American College, Madurai, Tamilnadan, India.

shaya7720@gmail.com

Abstract:

Multi-Agent Systems consist of autonomous intelligent agents that operate in the same environment to achieve their own goals or common goals. These agents can work without any human help, as they have a type of intelligence that allows them to make decisions and learn from the information they receive. However Multi-Agent Systems have to deal with some security problems, such as agents that are not authorized data getting out and bad things happening that can hurt the whole system. To make sure everything is safe agents need to be checked to see who they really are before they can get in, and this is done using keys and certificates. Over time the system also keeps an eye on how agents behave and what other agents think of them to see if they can be trusted. When agents talk to each other, they use a way of sending messages that keeps their information secret so it does not get out.This paper is talking about how to make Multi-Agent Systems really secure by using codes and keys. It suggests using things like SHA-256 hashing to store data digital signatures to verify messages are authentic and special tokens to control access. It even talks about using blockchain which is a way of storing records that can't be changed to keep track of trust relationships.

Keyword: Multi-Agent Systems, Agentic AI, Autonomous Agents, Digital Signatures, Privacy-Preserving Communication, Distributed Systems

1. Introduction

Multi-Agent Systems can be used in a variety of applications such as resource allocation in hospitals fraud detection in banks and traffic management in smart cities. This way of doing things has advantages, such as being able to handle a lot of agents keeping everything safe from attacks and being reliable. In the future the plan is to look into using intelligence to make trust models that can find problems in real time. It will also look at how to defend against threats that might arise as computers become even more powerful . One major area of research in artificial intelligence is Multi-Agent Systems . These are like environments where many autonomous agents live and work together to achieve their goals . Make their own decisions. What is special about Multi-Agent Systems is that they are made up of parts that work together not just one big system. This means that the agents can do things together that they could not do on their own and they can even come up with ways of doing things that are better than what any one agent could do by itself. Multi-Agent Systems are spread out over a network, which means that the agents can be, in places and still work together. Each agent can work on its own without being told what to do by a human. They can even work together to solve problems that are too hard for one agent to solve.

MAS applications cover critical societal infrastructure. In healthcare, networks of AI agents collaborate to diagnose disease, plan treatments and analyze drug interactions across geographically distributed hospital systems. In smart cities, MAS handles emergency response, energy distribution and traffic flow in real time.

Security threats to distributed AI systems come in many forms. In "agent impersonation" attacks, malicious actors impersonate trusted agents to intercept communications or influence judgments. In Sybil attacks, multiple fictitious agent identities are introduced to increase influence over group decisions. Whitewashing enables malicious agents to re-enter a network as apparently trustworthy newcomers, despite having compromised identities. Inter-agent communications in transit are intercepted and changed in man-in-the-middle attacks.

2. Literature Review

The academic community has paid a lot of attention to MAS security in the past two decades. The theoretical foundations of autonomous agents and their properties were established by Wooldridge and Jennings [1]. Shoham and Leyton-Brown extended this work to the study of multi-agent interactions and game-theoretic models of cooperation.

The MAS authentication mechanisms have evolved considerably from early password-based schemes to cryptographic solutions. The Public Key Infrastructure (PKI) offers a hierarchical trust model where the agent identities are vouched for by Certificate Authorities (CAs) using digitally signed certificates. Kambourakis showed that PKI can secure agent communications but noted the administrative burden of certificate lifecycle management in large-scale deployments. Using digital signatures which are derived from asymmetric cryptographic algorithms Agents can prove the authenticity and non-repudiation of messages

There has been much research on trust management in MAS using behavioural and reputation-based models. Kamvar et al. proposed the EigenTrust algorithm that adapts PageRank principles to compute global trust values for agents in peer-to-peer systems based on the collective opinions of the interacting peers. EigenTrust is effective but still susceptible to collusion attacks. Behavior-based trust models track the actions of agents over time to detect unusual patterns that may indicate compromise or malicious intent. Blockchain technology has recently been investigated as a decentralized trust substrate. Dorri et al. proposed blockchain-based IoT security frameworks, which have access to immutable audit trails and resistance to single points of failure. These frameworks can be used directly for MAS trust ledgers.

Despite this corpus of work, however, important research gaps remain. The integration of privacy, trust and authentication mechanisms into a single framework is not well STUDIED. The computational overhead of cryptographic operations in resource-constrained MAS environments is still not well understood. Furthermore, there has been no thorough evaluation of the robustness of existing trust models against advanced adversarial techniques such as coordinated whitewashing campaigns and adaptive Sybil attacks. The present study directly addresses these gaps.

3. Methods Of Research

In this work we employ a design-based research (DBR) approach that combines theoretical analysis, system design and iterative evaluation to produce theoretical insights and useful artifacts. The DBR approach is a natural fit for interdisciplinary work at the intersection of distributed systems, cybersecurity, and artificial intelligence, due to its ability to rigorously evaluate design choices in real-world settings.

The main contribution of this work is the Three-Pillar Security Framework that considers authentication, trust management and privacy protection as interrelated security properties that mutually reinforce each other. Each pillar was developed by systematically reviewing existing mechanisms, identifying gaps, and selecting or synthesizing mechanisms that fill those gaps.

These are: (1) Authentication, i.e., each agent must be able to authenticate any other agent it communicates with; (2) Trust Management, i.e., dynamically evaluating the trustworthiness of an agent based on its past behavior and reputation among its peers; and (3) Privacy Protection, i.e., preventing unauthorized disclosure of sensitive information exchanged between agents. In terms of evaluation methodology, we use comparative analysis along four dimensions. Scalability: how the mechanism behaves when the number of agents increases from tens to thousands? Resistance against attacks: the research uses simulation-based evaluation with agent-based modeling environments, complemented by theoretical complexity analysis. The case studies in healthcare, smart cities, and autonomous vehicles provide the necessary context for grounded evaluation.

4. System Architecture

The security architecture of MAS presented in this paper is five-layered. Each security concern is encapsulated into one interoperable layer. This allows for modular deployments where any layer can be updated or replaced without affecting the whole system. It also allows for a clear division of security responsibilities.

Agent Layer: Agents in this layer are responsible to initiate the authentication handshake, report behavioral anomalies to the trust layer and enforce local access control decisions.

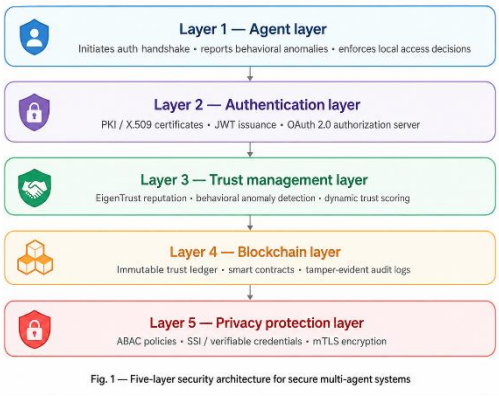


Fig. 1 — Five-layer security architecture for secure multi-agent systems

Authentication Layer: This layer is on top of the agent layer and is responsible for issuing, validating and revoking the agent identities. This includes the Public Key Infrastructure hierarchy with root and intermediate Certificate Authority, JSON Web Token token issuance service and OAuth 2.0 authorization server.

Trust Management Layer: It handles dynamic trust score management for each agent in the system, aggregates peer feedback with Eigen Trust-based algorithm, monitors behavior streams for anomalous pattern detection using statistical anomaly detection and publishes trust scores periodically to the blockchain layer for tamper-evident recording.

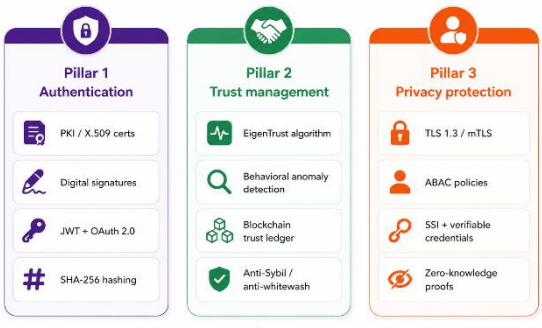
Blockchain Layer: It provides a distributed and immutable ledger for trust score records, identity credential hashes and audit logs. Smart contracts allow the automatic updating of trust scores and the enforcement of access control policies without a trusted central authority, thus improving the transparency and accountability of the MAS

Privacy Protection Layer: This layer enforces data access policies using Attribute-Based Access Control (ABAC) rules and Self-Sovereign Identity (SSI) credential management for selective disclosure. It intercepts all inter-agent data exchanges to allow data sharing only with those agents with the required attribute credentials and enforces Transport Layer Security (TLS) on all communications.

Secure communication. The initiating agent presents its JWT and certificate to the authentication layer, which verifies the credentials, and queries the trust layer for the trust score of the target agent. If the score is above a configurable threshold, the privacy layer checks the attribute credentials of the initiating agent against the ABAC policy of the target resource, and only if this authorization is passed does the communication continue, encrypted by mutual TLS.

Smart contracts automatically update trust scores and enforce access control policies without any trusted central authority, resulting in increased transparency and accountability of the MAS.

5. Security Framework Design



5.1 Pillar 1: Authentication

The authentication pillar considers verified agent identities as the foundation on which the MAS security stands. Public Key Infrastructure (PKI) is the situation in which each agent is given an X.509 certificate signed by a trusted CA, which binds a public key to the identity of the agent and can be verified by any other agent through a signature from the CA. All the inter-agent messages are digitally signed with the agent's private key to provide authentication and non-repudiation, i.e. the recipient of the message can check the identity of the sender and can prove that the message was not modified during transit. Token-based authentication with JWT and OAuth 2.0, which is stateless and scalable, complements certificate-based identity for runtime authentication in dynamic environments.

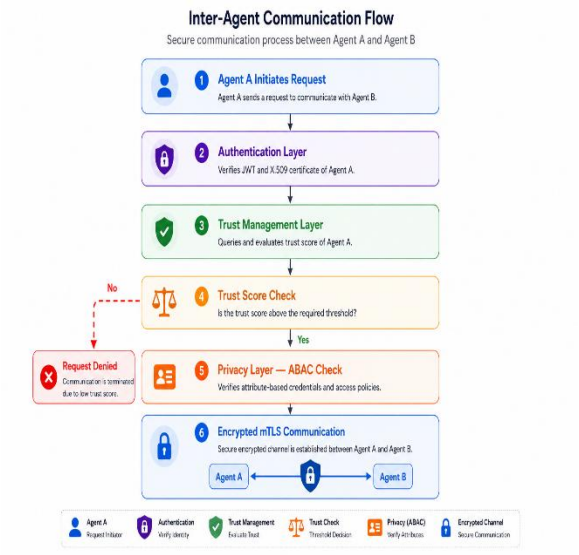
An agent authenticates an authorization server with its certificate credentials and is provided with a signed JWT containing claims about its identity and the scopes it is authorized to have. This token is then carried in subsequent communications, and a certificate revocation does not need to be verified each time. While PKI supports certificates, widely known limitations include the cumbersome complexity of revocation and the administrative overhead in managing large scales of CA operation – substantial overhead in storage, computation, and communications.

5.2 Pillar 2: Trust Management

This pillar of trust management dynamically assesses agent reliability in context. At the core of the reputation mechanism is the Eigen Trust algorithm, where each agent maintains a local trust vector for its peer assessments; these vectors are aggregated at the network level through normalized matrix computation. Behavioral anomaly detection (BAD) is a complementary approach to reputation-based trust, whereby the actions of the agents are monitored in real-time. For each agent, normal behaviour baselines are statistically determined, and deviations from these baselines exceeding configured thresholds result in deductions to trust scores and the issuance of alert messages. This is particularly true for insider attacks where a legitimate agent can be compromised after successful authentication. Each trust score update is recorded on the blockchain trust ledger through smart contract transactions that provide an immutable audit trail. Sybil attacks are prevented by requiring proof-of-work or proof-of-stake during registration of new agents and whitewashing attacks are prevented by new identities cryptographically committing to at least a part of the previous identity's trust history on the blockchain.

5.3 Pillar 3 : privacy protection

The privacy protection pillar guarantees that sensitive information is only exchanged between agents that have a legitimate need and are authorized to access it. All communications between the agents are encrypted using Transport Layer Security (TLS) 1.3 and Mutual TLS (MTLS). This prevents eavesdropping, man-in-the-middle attacks. MTLS also requires the receiving agent to authenticate to the sending agent, thus providing a mutual identity verification on the transport layer. ACLs provide a coarse-grained way to control access to resources. Fine-grained, policy-driven authorization based on agent attributes like role, department, clearance level and location is possible using Attribute-Based Access Control (ABAC). ABAC policies are defined as logical rules that are evaluated against the attributes of the requesting agent, allowing for fine-grained access decisions that are sensitive to the context of each interaction. The W3C Decentralized Identifier (DID) and Verifiable Credential (VC) standards enable the creation of Self-Sovereign Identity frameworks where agents can manage their own identity claims without reliance on centralized identity providers. Selective disclosure is a zero knowledge proof technique where an agent can prove that they possess a credential attribute without revealing the credential itself, thus minimizing information exposure significantly.



6. Results

The proposed Three-Pillar Security Framework is evaluated against existing single-mechanism approaches across two dimensions: attack resistance and known limitations. The analysis is theoretical, based on the design properties of each mechanism.

6.1 Attack Resistance

Table 1 shows the comparison of the resistance to five common MAS attack types. Each single-mechanism approach leaves big gaps: PKI does not address Sybil or insider attacks, EigenTrust does not prevent impersonation, and blockchain alone cannot stop man-in-the-middle attacks. The integrated framework builds on complementary defences for strong resistance across all categories: certificate verification against impersonation, proof-of-work registration against Sybil attacks, cryptographic identity binding against whitewashing, mTLS against eavesdropping and behavioural anomaly detection against insider threats.

Table 1: Attack resistance comparison across security mechanisms

Attack type	PKI-only	EigenTrust-only	Blockchain-only	Three-Pillar Framework
Agent impersonation	Partial	None	None	Strong
Sybil attack	None	Moderate	Moderate	Strong
Whitewashing	None	Weak	Moderate	Strong
Man-in-the-middle	Partial	None	None	Strong
Insider attack	None	Moderate	Weak	Strong

6.2 Limitations

While the framework encompasses much, it has three trade-offs that future empirical evaluation should tackle, as summarized in Table 2.

Table 2: Known limitations of the Three-Pillar Framework

Limitation	Description
Computational overhead	Cryptographic operations and blockchain consensus introduce latency in time-critical deployments.
Blockchain dependency	Consensus latency makes the framework unsuitable for disconnected or low-bandwidth MAS environments.
Eigen Trust boundary	Trust convergence degrades when the proportion of malicious agents approaches 50%.

These limitations set the boundaries for the applicability of the framework and outline clear directions for future work, such as lightweight cryptographic alternatives, hybrid ledger architectures, and adaptive trust bootstrapping for adversarial environments.

7. Conclusion

Multi-Agent Systems are gaining more and more importance for critical infrastructure. The distributed and autonomous nature of these systems makes them more vulnerable to security threats that cannot be fully protected by any single system. The current paper has identified a major gap in the current literature with the absence of a unified security framework in a single architectural model to address agent authentication, dynamic trust assessment, and privacy-preserving data exchange.

The proposed Three-Pillar Security Framework fills this gap by integrating PKI, JWT, OAuth 2.0, EigenTrust-based reputation systems, behavioral anomaly detection, blockchain trust ledgers, ABAC, and Self-Sovereign Identity in a succinct five-layer framework. Each pillar supports the others: trusted agents provide trust scores that inform access decisions which are enforced by the privacy layer.

This interdependence results in a total security posture that is significantly stronger than the sum of its individual parts.

This framework's real-world relevance is demonstrated in the healthcare, smart city and autonomous transportation sectors analyzed in this study where agent compromise results in outcomes ranging from incorrect medical dosing to systemic infrastructure failures. The framework works well with how things operate in these high-stakes environments, baking in security from the start rather than adding it at the end.

Verifiable identity, dynamic trust, and privacy by design will be the most critical components of trustworthy multi-agent ecosystems, as AI systems become more and more decentralized and autonomous.

Decentralized AI safety is a technical challenge and a necessary step to the safe deployment of intelligent systems in society.

References

[1] *Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems : (CBMS 2002) : 4-7 June, 2002, Maribor, Slovenia*. Los Alamitos, Calif.: IEEE Computer Society,.

[2] H. Aziz, "Multiagent systems," *ACM SIGACT News*, vol. 41, no. 1, p. 34, 2010, doi: https://doi.org/10.1145/1753171.1753181.

[3] S. Hota, *Emerging Trends in Artificial Intelligence Based IoT: Techniques, Applications and Security*. Bentham Science Publishers, 2025.

[4] Sandro Etalle, S. Marsh, and SpringerLink (Online Service, *Trust Management : Proceedings of IFIPTM 2007: Joint iTrust and PST Conferences on Privacy, Trust Management and Security, July 30-August 2, 2007, New Brunswick, Canada*. New York, Ny: Springer Us, 2007.

[5] M. Wooldridge, *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2002.

[6] A. Birukou, E. Blanzieri, and P. Giorgini, "Implicit: a multi-agent recommendation system for web search," *Autonomous Agents and Multi-Agent Systems*, vol. 24, no. 1, pp. 141–174, Aug. 2010, doi: https://doi.org/10.1007/s10458-010-9148-z.

[7] F. Pauravi and A. Latif, "Secure Multi Agent Information Filtering," *International Journal of Computer Theory and Engineering*, vol. 6, no. 3, pp. 240–246, 2014, doi: https://doi.org/10.7763/ijcte.2014.v6.869.

[8] J. Babu and K. G, "A Secure Communication for a Reputation Management Model in Multi- agent System," *International Journal of Computer Applications*, vol. 84, no. 2, pp. 19–23, Dec. 2013, doi: https://doi.org/10.5120/14549-26