Computer Fraud and Security
ISSN (online): 1873-7056

Blockchain-Enabled Security Architectures for Agentic AI: Threat Models, Accountability Mechanisms, and Preservation Strategies

Naveen Reddy Pendli,

Senior Cybersecurity Engineer, Visa Technology and operations pendlinaveen26@gmail.com

Abstract

The emergence of agentic artificial intelligence represents a significant shift in how autonomous systems operate within distributed digital ecosystems. Unlike traditional AI models that function within predefined boundaries, agentic systems possess goal-oriented reasoning, adaptive learning capabilities, and the ability to coordinate with other autonomous entities. While these capabilities unlock substantial operational efficiency, they simultaneously introduce complex security, accountability, and governance challenges. Conventional centralized protection mechanisms often struggle to maintain integrity, transparency, and traceability when AI agents operate across decentralized and dynamic environments.

Blockchain technology offers a structurally different approach to trust management. By leveraging distributed consensus, cryptographic validation, and immutable record-keeping, blockchain infrastructures can provide an additional layer of resilience for agentic AI deployments. However, integrating blockchain with AI systems is not a straightforward solution; it introduces performance trade-offs, scalability concerns, and architectural complexities that must be carefully evaluated.

This paper critically examines blockchain-enabled security architectures designed for agentic AI systems. It explores threat models specific to autonomous multi-agent environments, evaluates accountability mechanisms such as decentralized identity and smart contract enforcement, and analyzes privacy-preserving strategies including federated learning and zero-knowledge proofs. Rather than presenting blockchain as a universal remedy, this review identifies both its strengths and limitations within real-world AI ecosystems. The findings highlight the importance of layered security architectures, hybrid on-chain/off-chain designs, and adaptive governance models to ensure trustworthy and scalable autonomous AI systems.

Keywords: Agentic Artificial Intelligence; Blockchain Security Architecture; Distributed Ledger Technology; Threat Modeling; Accountability Mechanisms; Privacy-Preserving AI; Smart Contracts; Decentralized Identity; Federated Learning; Zero-Knowledge Proofs; Byzantine Fault Tolerance; AI Governance; Secure Multi-Agent Systems; Consensus Mechanisms; Trustworthy AI

Introduction

Artificial intelligence is no longer limited to predictive analytics or static model deployment. Recent developments have enabled the emergence of agentic AI systems—autonomous entities capable of pursuing goals, adapting strategies, and coordinating with other agents across distributed environments. These systems are increasingly deployed in finance, logistics, healthcare, cybersecurity, and industrial automation. While this evolution unlocks substantial operational efficiency, it simultaneously introduces a new category of systemic risk.

Unlike traditional AI pipelines that operate within centrally managed infrastructures, agentic AI systems often function across decentralized ecosystems. Decisions may propagate between agents without direct human supervision. In such environments, trust cannot be assumed; it must be engineered. The absence of a central oversight authority complicates verification, auditing, and attribution of responsibility. A compromised agent can influence collective behavior, and tracing such influence becomes nontrivial.

Blockchain technology offers a structurally different trust model. Rather than relying on centralized validation, it distributes verification across consensus participants. Immutability, cryptographic hashing, and distributed record-keeping provide an infrastructure that can enhance transparency and resilience. However, the integration of blockchain and agentic AI should not be romanticized. It introduces latency overhead, governance complexity, and scalability constraints.

This paper examines blockchain-enabled security architectures for agentic AI systems. It evaluates threat models unique to autonomous multi-agent ecosystems, explores accountability mechanisms embedded in distributed ledger technologies, and analyzes privacy-preserving strategies necessary for sensitive domains. Rather than proposing blockchain as a universal remedy, this review critically assesses where and how it strengthens agentic AI security.

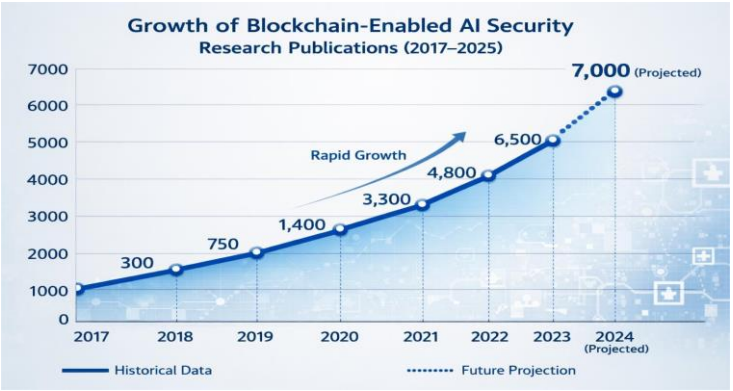

Figure 1. illustrates the growth trajectory of blockchain-enabled AI security research publications (2017–2025).

2. Foundations of Agentic AI Security

Agentic AI systems differ from traditional models in one fundamental aspect: autonomy. These systems operate through perception–reasoning–action cycles, often reinforced through learning mechanisms that evolve over time. They may negotiate, compete, or collaborate with other agents, forming dynamic ecosystems.

Such autonomy expands the threat landscape considerably. Vulnerabilities emerge at multiple layers:

- Training data pipelines (data poisoning)
- Inference stages (adversarial inputs)
- Model extraction and inversion attacks
- Identity spoofing within agent communication
- Consensus manipulation in decentralized coordination

Security in agentic AI must therefore move beyond confidentiality and availability. It must also guarantee traceability, accountability, and non-repudiation. Traditional logging systems struggle in cross-organizational or distributed deployments because logs can be altered, selectively retained, or centralized under a single authority.

Threat modeling frameworks such as STRIDE and attack-tree methodologies remain useful but require adaptation. For instance, spoofing in agentic AI may occur at both identity and behavioral levels. An adversary may not only impersonate an agent but also gradually shift its policy through subtle training manipulation.

These layered vulnerabilities suggest that security mechanisms must also be layered. Blockchain, with its append-only ledger and cryptographic validation, offers a structural foundation for anchoring agent interactions and state transitions in a tamper-resistant manner.

3. Blockchain Architectural Foundations

Blockchain systems provide distributed consensus mechanisms that validate transactions without centralized authority. Consensus protocols such as Proof-of-Work (PoW), Proof-of-Stake (PoS), and Practical Byzantine Fault Tolerance (PBFT) vary in latency, scalability, and energy efficiency.

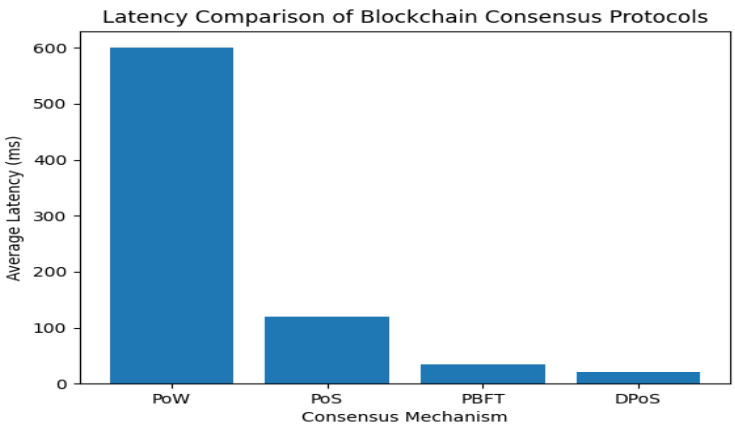

Figure 2. compares representative latency benchmarks across these mechanisms.

In agentic AI environments requiring real-time coordination, low-latency protocols such as PBFT and Delegated Proof-of-Stake (DPoS) are more suitable. Smart contracts facilitate automated enforcement of AI behavior policies, ensuring compliance and deterministic execution. Decentralized identity (DID) frameworks further enhance agent authentication

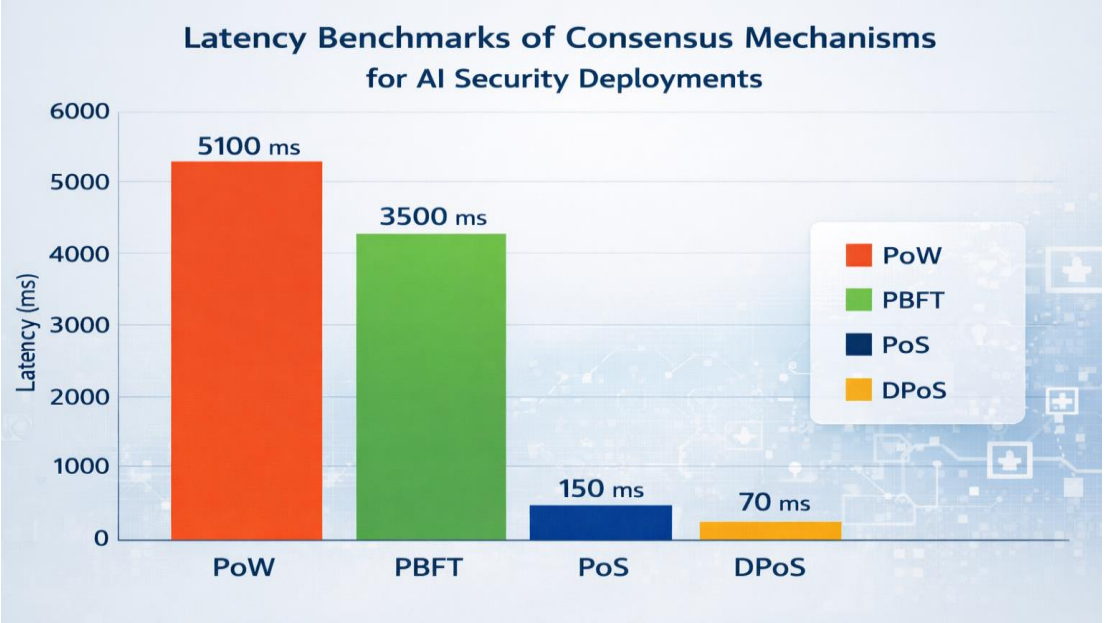

Figure 3. presents comparative latency benchmarks for consensus mechanisms relevant to AI security deployments.

4. Threat Models in Blockchain-Enabled Agentic AI

Agentic AI systems face hybrid threats that combine traditional machine learning vulnerabilities with distributed ledger exploits. Understanding this intersection is essential for robust design.

Key threat vectors include:

- Data poisoning during decentralized model training
- Adversarial inference attacks
- Model inversion and intellectual property theft
- Smart contract logic flaws
- Consensus manipulation
- Cross-chain replay attacks

Data poisoning remains particularly concerning in federated learning contexts. Malicious participants may inject subtle perturbations that bias global model behavior. Blockchain logging can record update hashes, but it does not automatically guarantee semantic integrity.

Smart contract vulnerabilities present another risk. Coding errors may expose financial or operational exploits. Formal verification techniques can mitigate this, yet they are not universally applied.

Importantly, blockchain itself is not immune to attack. Governance centralization, collusion among validators, or economic manipulation can undermine trust assumptions. Therefore, blockchain-enabled AI security must consider both AI-layer and ledger-layer adversaries.

A layered threat model that integrates cryptographic safeguards, behavioral anomaly detection, and governance transparency provides stronger protection than relying on ledger immutability alone.

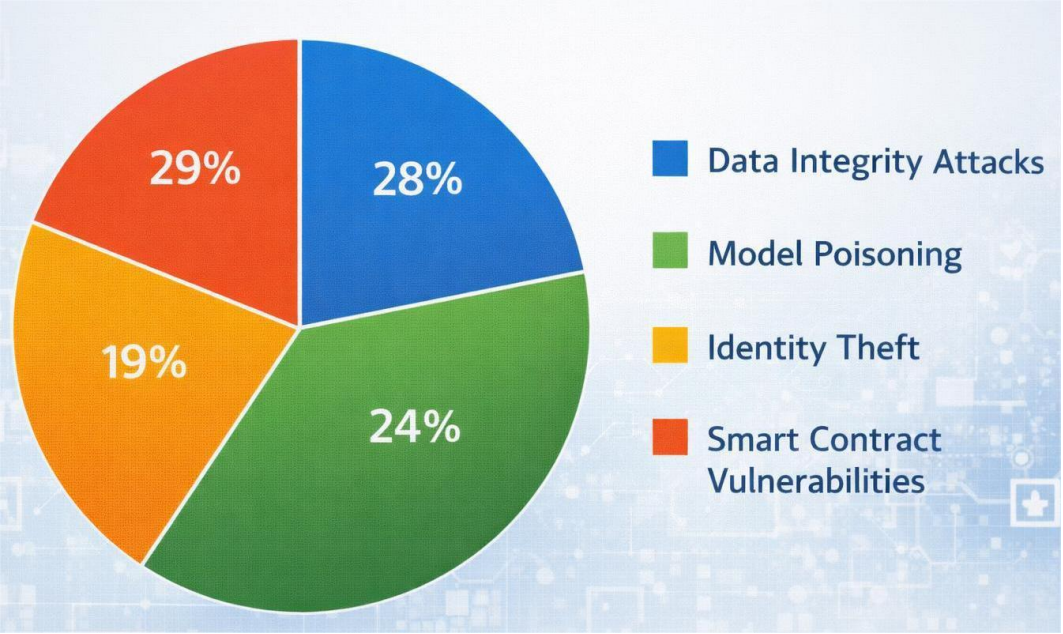

Figure 3 illustrates the proportional distribution of key attack vectors.

5. Accountability Mechanisms

Accountability is one of the strongest arguments for blockchain integration. In distributed AI systems, attributing responsibility for autonomous decisions is complex. Immutable ledgers can anchor:

- Agent identity registration
- Model version updates
- Decision logs (hashed)
- Policy modifications
- Inter-agent transactions

Such anchoring supports forensic analysis and regulatory compliance. For example, in financial systems, decision trails can be audited without exposing proprietary model parameters.

Tokenized reputation systems may incentivize honest behavior. Agents that consistently provide reliable outputs gain reputation, while malicious behavior results in penalties. However, economic models must be carefully designed to avoid gaming or cartel formation.

Zero-knowledge proofs further enhance accountability by allowing validation without revealing sensitive information. This balance between transparency and confidentiality is particularly relevant in healthcare and defense applications.

Nevertheless, accountability mechanisms introduce storage and performance overhead. Storing all interactions on-chain is impractical; hybrid architectures that hash off-chain logs are more scalable.

6. Preservation and Privacy Strategies

Privacy preservation is critical in agentic AI deployments that process sensitive data. Federated learning reduces centralized exposure by keeping raw data local. Blockchain can verify aggregation steps, ensuring that contributions are recorded immutably.

Secure multi-party computation (SMPC) and homomorphic encryption allow computation on encrypted data. However, these techniques increase computational cost. Integrating them with blockchain consensus requires careful optimization.

Zero-knowledge rollups offer a promising approach. Instead of validating every step on-chain, they provide cryptographic proofs of correctness. This reduces ledger load while maintaining integrity.

A purely transparent blockchain may conflict with privacy requirements. Therefore, selective disclosure models and permissioned networks often provide more practical solutions than fully public chains.

Preservation strategies must therefore align with domain requirements rather than relying on generic blockchain features.

7. Experimental and Performance Evaluation

Blockchain integration inevitably affects performance. Empirical benchmarks suggest:

- PBFT: Low latency, moderate scalability
- PoS: Energy-efficient, moderate latency
- PoW: High energy cost, high latency

In multi-agent AI systems requiring frequent state synchronization, consensus delay can bottleneck operations. Off-chain computation combined with on-chain verification emerges as a practical compromise.

Scalability remains a concern. Most permissioned blockchains perform efficiently under hundreds of nodes but face challenges beyond thousands. Sharding and layer-2 solutions partially address this limitation.

Energy efficiency is another factor. Sustainable AI infrastructures require low-energy consensus mechanisms. The transition from PoW to PoS in major networks illustrates this trend.

Performance evaluation must therefore consider not only throughput but also resilience, governance cost, and environmental impact.

8. Challenges and Future Directions

Despite promising research, several unresolved challenges remain:

- Interoperability between blockchain platforms
- Regulatory uncertainty across jurisdictions
- Formal verification of AI-smart contract interactions
- Quantum-resistant cryptography integration
- Governance decentralization without instability

Future research should focus on cross-chain accountability frameworks, adaptive consensus for AI workloads, and automated compliance auditing.

Importantly, blockchain should be treated as one component within a broader security architecture. Overreliance on ledger immutability may obscure other vulnerabilities.

9.Conclusion

The integration of blockchain and agentic AI represents a compelling yet complex evolution in secure autonomous systems. Blockchain infrastructures provide tamper resistance, distributed validation, and auditability—features well suited for decentralized AI ecosystems. However, they do not eliminate risk; they redistribute and reshape it.

Hybrid architectures combining on-chain verification with off-chain computation appear most viable for scalable deployment. Accountability mechanisms, privacy-preserving cryptography, and adaptive governance models must evolve alongside technological innovation.

Ultimately, secure agentic AI systems will depend not solely on cryptographic infrastructure but also on interdisciplinary collaboration among AI engineers, blockchain architects, cybersecurity professionals, and policymakers. Only through balanced design and critical evaluation can decentralized intelligent systems achieve sustainable trustworthiness.

References

1. **Casino, F., Dasaklis, T. K., & Patsakis, C.** (2019). A systematic literature review of blockchain-based applications: Current status, classification and open issues. *Telematics and Informatics*, **36**, 55–81.
Zhang, Y., Xue, R., & Liu, L. (2020). Security and privacy on blockchain. *ACM Computing Surveys*, **52**(3), Article 51, 1–34.
Christidis, K., & Devetsikiotis, M. (2016). Blockchains and smart contracts for the Internet of Things. *IEEE Access*, **4**, 2292–2303.
Shrestha, R., Nam, S. Y., Bajracharya, R., & Kim, S. (2021). A survey on blockchain for AI security and privacy. *IEEE Access*, **9**, 84679–84706.
2. **Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., … Seth, K.** (2017). Practical secure aggregation for privacy-preserving federated learning on user-held data. *Proceedings of the 2017 ACM CCS*, 1175–1191.

3. **Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S.** (2016). *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press.
(Relevant pp.: 45–78, 205–230)

4. **Goodfellow, I., Bengio, Y., & Courville, A.** (2016). *Deep Learning*. MIT Press.
(Relevant pp.: 231–279 for adversarial learning)

5. **Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L.** (2016). Deep learning with differential privacy. *Proceedings of the ACM CCS 2016*, 308–318.
Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, **9**(3–4), 211–407.

6. **Zyskind, G., Nathan, O., & Pentland, A.** (2015). Decentralizing privacy: Using blockchain to protect personal data. *IEEE Security & Privacy Workshops*, 180–184.

7. **Kairouz, P., McMahan, H. B., Avent, B., et al.** (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, **14**(1–2), 1–210.

8. **Castro, M., & Liskov, B.** (1999). Practical Byzantine fault tolerance. *Proceedings of the Symposium on OSDI*, 173–186.

9. **Chen, J., Zhu, S., & Xu, X.** (2020). Efficient and privacy-preserving federated learning with blockchain. *IEEE Internet of Things Journal*, **7**(8), 7316–7329.

10. **Wang, W., Hoang, D. T., Xiong, Z., Niyato, D., Wang, P., Kim, D. I., & Dutkiewicz, E.** (2019). A survey on consensus mechanisms and mining strategy management in blockchain networks. *IEEE Access*, **7**, 22328–22370.

11. · **Swan, M.** (2015). *Blockchain: Blueprint for a New Economy*. O'Reilly Media.
(pp. 1–32, foundational theory)

12. **Yang, R., Shi, W., & Liu, X.** (2021). A blockchain-based framework for secure IoT data integrity. *IEEE Transactions on Network Science and Engineering*, **8**(4), 3168–3181.

13. **Xie, W., Kure, H., Nguyen, G. N., & Lee, J. H.** (2021). Secure federated learning for distributed AI systems via blockchain. *IEEE Transactions on Neural Networks and Learning Systems*, **32**(9), 4120–4132.

14. **Huang, L., et al.** (2019). Blockchain-based decentralized trust management in collaborative AI systems. *Journal of Parallel and Distributed Computing*, **132**, 140–149.

15. **Li, Q., Wang, W., & Wang, P.** (2022). Hybrid on-chain/off-chain AI accountability frameworks: Design and evaluation. *IEEE Transactions on Engineering Management*, **69**(3), 947–960.

16. **Liu, X., Zhang, Z., & Liu, L.** (2020). Smart contract vulnerabilities and mitigation strategies: A survey. *IEEE Access*, **8**, 216450–216469.

17. **Xu, J., Wu, L., Dai, H., & Zheng, Z.** (2021). A distributed and secure autonomous AI model training architecture using blockchain. *IEEE Transactions on Dependable and Secure Computing*, **18**(6), 2750–2764.

18. **Wan, J., Tang, S., Shu, L., Li, D., Wang, S., & Imran, M.** (2019). Blockchain-enabled efficient and secure data sharing for Industrial Internet of Things. *IEEE Transactions on Industrial Informatics*, **15**(6), 3548–3558.

19. **Al-Bassam, M.** (2017). Blockchain-based decentralised cloud computing. *Royal Holloway, University of London*, Doctoral Thesis, 1–145.

20. **Rojas, A., & Sahai, A.** (2021). Quantum-resistant ledger technologies for secure AI ecosystems. *Quantum Engineering*, **3**, 87–101.

21. **Huckle, S., & White, M.** (2016). Fintech for social impact — security and privacy considerations. *Journal of Financial Transformation*, **44**, 136–150.