Enhancing Model Accountability via Automated Lineage Tracking and Cryptographic Proofs of Decision Integrity in Financial AI

Jeremy Hollis
Department of Financial Engineering, Lehigh University
j.hollis@lehigh.edu

Alan Kingsley
School of Computing and Information, University of Pittsburgh
alan.kingsley@pitt.edu

Abstract

The proliferation of high-frequency algorithmic decision-making within global financial markets has outpaced existing regulatory frameworks, creating an accountability gap that threatens systemic stability and public trust. As financial institutions increasingly deploy opaque artificial intelligence models for credit scoring, risk assessment, and autonomous trading, the inability to reconstruct the exact state of a system at the moment of a specific decision poses significant legal and operational risks. This paper proposes a comprehensive socio-technical architecture designed to enhance model accountability through two primary technological pillars: automated lineage tracking and cryptographic proofs of decision integrity. Automated lineage tracking ensures a continuous, immutable record of data provenance, hyperparameter configurations, and model versions, allowing for granular historical reconstruction. Concurrently, the integration of cryptographic primitives, such as zero-knowledge proofs and secure hardware enclaves, provides a mechanism for verifying that a specific decision was generated by an authorized version of a model without necessitating the exposure of proprietary algorithmic logic. We analyze the structural trade-offs inherent in deploying such a high-integrity infrastructure, specifically the tensions between computational latency, storage overhead, and regulatory transparency. The discussion extends to the governance implications of these technologies, emphasizing how they facilitate "accountability by design" and support compliance with emerging international standards. By examining the intersection of distributed systems, financial regulation, and machine learning, this research offers a robust framework for ensuring that autonomous financial agents remain tethered to human oversight and institutional responsibility, ultimately fostering a more resilient and fair financial ecosystem.

Keywords:
Financial AI, Model Accountability, Data Lineage, Cryptographic Verification, Algorithmic Governance, Systemic Risk

1. Introduction

The evolution of financial services over the past decade has been defined by a decisive shift toward automation, characterized by the deployment of increasingly complex machine learning architectures. These systems now manage trillions of dollars in assets, determine the creditworthiness of millions of individuals, and execute trades at microsecond intervals. While the efficiency gains of financial artificial intelligence are undeniable, they have introduced a profound crisis of accountability. In the event of a market flash crash, a biased lending decision, or a localized systemic failure, identifying the root cause within a modern AI pipeline is often an exercise in futility. Traditional auditing methods, designed for human-readable logic and static rule-sets, are fundamentally unsuited for the dynamic, non-linear nature of deep learning and ensemble models.

The challenge of accountability in financial AI is exacerbated by the multi-staged nature of the machine learning lifecycle. A single decision is the culmination of a complex lineage involving raw data ingestion, feature engineering, model training, validation, and real-time inference. Without a rigorous mechanism to track this lineage, it becomes impossible to determine whether a faulty output was the result of corrupted input data, a misconfigured training parameter, or an adversarial perturbation during deployment. Moreover, the proprietary nature of financial algorithms creates a transparency paradox: regulators require proof of integrity, yet institutions must protect their intellectual property. To resolve this, a new paradigm of accountability is required—one that utilizes automated metadata management and cryptographic verification to provide irrefutable evidence of a system's state and logic without compromising competitive advantages.

2. The Infrastructure of Automated Lineage Tracking

At the core of a robust accountability framework lies the capacity for comprehensive data and model lineage tracking. Lineage tracking refers to the systematic documentation of the entire history of a digital artifact, from its origin through every transformation it undergoes within a system. In a financial AI context, this necessitates an automated infrastructure that captures metadata at every stage of the pipeline. This includes the exact version of the datasets used for training, the specific software dependencies of the environment, the hyperparameter settings, and the unique identifier of the model weights. By creating an immutable graph of these relationships, organizations can perform "post-mortem" analysis with surgical precision, reconstructing the exact environment that produced a specific financial outcome [12].

The implementation of such an infrastructure requires a departure from ad-hoc logging toward a centralized metadata repository integrated into the Continuous Integration and Continuous Deployment (CI/CD) pipeline. This metadata layer acts as the system’s "black box" flight recorder. When a credit application is rejected by an automated agent, the lineage tracker can instantly link that specific decision to the version of the model that was live at that millisecond, the training data that influenced the model's weights, and the validation tests that the model passed prior to deployment. This level of granularity is essential for addressing issues of fairness and bias. For example, if a model is found to be discriminating against a

specific demographic, lineage tracking allows researchers to trace the bias back to specific skewed data sources or problematic feature engineering steps [32]. This transparency is not merely a technical luxury but a foundational requirement for sustainable AI governance.

3. Cryptographic Proofs of Decision Integrity

While lineage tracking provides a historical record, it does not inherently prove that the record is untampered or that the model executed as intended. Cryptographic proofs of decision integrity provide the second necessary layer of accountability. In high-stakes financial environments, it is possible for malicious actors—either internal or external—to alter a model's behavior or forge logs to hide illicit activity. To prevent this, we propose the integration of cryptographic primitives that bind the output of a model to its specific version and the input data. One promising approach involves the use of zero-knowledge proofs (ZKPs), which allow an institution to prove to a regulator that a specific decision was computed correctly according to a registered model without revealing the internal parameters of that model [8].

Furthermore, the deployment of models within Trusted Execution Environments (TEEs) or secure hardware enclaves ensures that the inference process is isolated from the host operating system. This prevents the "man-in-the-middle" modification of decisions or data. By signing the output of the model within the enclave, the system generates a cryptographic certificate of integrity. This certificate serves as a digital seal, guaranteeing that the decision was produced by the authorized, audited version of the algorithm. This architecture addresses the structural trade-off between privacy and transparency. It allows for a "trust-but-verify" relationship between financial institutions and regulatory bodies, where the integrity of the process is mathematically guaranteed even when the underlying code remains confidential [15]. Such mechanisms are particularly vital in decentralized finance (DeFi) and cross-border transactions where traditional jurisdictional oversight is often fragmented or weak [22].

4. Structural Trade-offs in High-Integrity Systems

The transition toward high-integrity financial AI infrastructures involves significant structural trade-offs, primarily concerning performance, cost, and complexity. The introduction of automated lineage tracking and cryptographic signing adds non-trivial computational overhead to the inference pipeline. In the domain of high-frequency trading, where every microsecond translates into significant financial gain or loss, the latency introduced by cryptographic operations can be a deterrent to adoption. Engineering teams must therefore balance the depth of the audit trail with the real-time requirements of the application. This often leads to the implementation of tiered accountability levels: real-time trading systems may utilize lightweight signing and asynchronous logging, while less time-sensitive applications like mortgage processing or long-term risk assessment employ full cryptographic verification and extensive lineage documentation [19].

Storage and data management present another significant hurdle. An automated lineage system generating metadata for every micro-decision can quickly produce petabytes of logs, creating a "data swamp" that is difficult to navigate and expensive to maintain. Ensuring the

sustainability of these systems requires sophisticated data lifecycle policies, where metadata is indexed, compressed, and eventually archived or pruned based on regulatory retention requirements. Moreover, the robustness of the cryptographic layer is only as strong as the underlying key management infrastructure. If the private keys used to sign model decisions are compromised, the entire accountability framework collapses. Thus, the deployment of these systems necessitates a concomitant investment in physical and digital security protocols, further increasing the operational complexity of the financial institution [27].

5. Governance, Fairness, and Policy Implications

The technological solutions proposed in this paper must be situated within a broader governance framework to be effective. "Accountability by design" implies that ethical and legal considerations are integrated into the technical architecture from the outset, rather than being treated as an afterthought. For regulators, the existence of automated lineage and cryptographic proofs simplifies the auditing process, moving from periodic, manual reviews toward continuous, automated oversight. This shift allows for the development of "regulatory APIs," where government agencies can query the integrity certificates and lineage graphs of financial institutions in real-time, significantly reducing the time required to detect and mitigate systemic risks or market abuses [4].

The implications for fairness and social equity are equally profound. In many instances, the "cultural gap" in data and model generation can lead to the marginalization of specific groups, a phenomenon that is often hidden behind the technical complexity of AI systems [22]. By mandating detailed lineage tracking, policy-makers can require institutions to prove that their models were trained on diverse and representative datasets. If an automated system produces a disparate impact, the lineage record provides a clear path for remediation. Furthermore, the use of cryptographic proofs can empower consumers. An individual whose loan was denied by an AI could be provided with a verifiable proof that the decision was made according to the bank’s stated policies, rather than an arbitrary or discriminatory impulse. This enhances the legitimacy of automated systems and helps to rebuild the public trust that has been eroded by the "black box" nature of contemporary finance [11].

6. Robustness and Adversarial Resilience in Financial AI

Financial systems are increasingly the targets of adversarial attacks designed to manipulate model outputs for profit. These attacks can take the form of "poisoning" training data to create backdoors in the model or craftily designed input perturbations that cause the model to misclassify risks. A system lacking lineage tracking is particularly vulnerable to such attacks, as the source of the corruption may go undetected for months. By implementing a high-integrity architecture, institutions gain a critical defensive tool: the ability to verify the integrity of every component in the machine learning stack. If a model begins to exhibit anomalous behavior, the lineage tracker can be used to isolate the exact moment the degradation began and identify the specific data batches or code changes that preceded it [3].

Robustness in this context also refers to the system’s ability to remain accountable during periods of high market volatility. During a market crisis, autonomous agents often encounter

"out-of-distribution" scenarios that were not present in their training data. In these moments, the logic of the model may break down, leading to irrational or cascading sell-offs. An accountable infrastructure ensures that these failures are documented and attributable. Cryptographic proofs can be used to trigger "kill-switches" or "fail-safe" modes: if a model cannot produce a proof that its current decision falls within pre-defined safety bounds, the system can automatically transition to a manual or conservative operating state. This dynamic resilience is essential for preventing the type of technical debt and systemic fragility that characterizes many current financial infrastructures [14].

7. Global Deployment and Standardization

As financial institutions operate across multiple jurisdictions, the lack of standardized accountability protocols remains a significant barrier to the global deployment of AI. A model that meets the accountability standards of the European Union may not satisfy the requirements of the United States or Asian markets. The integration of automated lineage and cryptographic proofs offers a path toward a "universal audit language." If the integrity of a model is verified by a mathematically-sound proof, that proof can be recognized and validated by any regulator, regardless of their location. This would facilitate the cross-border movement of algorithmic assets and reduce the compliance burden for multi-national corporations [25].

However, achieving this level of global standardization requires a coordinated effort between industry, academia, and government. We advocate for the development of open-source standards for metadata schemas and cryptographic verification protocols. By creating a shared infrastructure for accountability, the financial sector can avoid the fragmentation and "vendor lock-in" that often plagues high-tech industries. Moreover, a standardized approach ensures that smaller financial institutions, which may lack the resources to build proprietary accountability systems from scratch, can still participate in the AI-driven economy while maintaining the necessary level of oversight. The sustainability of the global financial system depends on our ability to create a transparent, verifiable, and equitable technological foundation that can support the next generation of autonomous finance [30].

8. Conclusion

The integration of artificial intelligence into the financial sector represents a permanent shift in how capital is allocated and risk is managed. To ensure that this shift does not come at the cost of systemic stability or social justice, we must bridge the accountability gap through rigorous technological and structural innovation. Automated lineage tracking and cryptographic proofs of decision integrity provide the necessary tools to transform "black box" algorithms into transparent, auditable, and responsible agents. By creating an immutable record of a system's evolution and a verifiable proof of its execution, we can ensure that autonomous financial systems remain under the effective control of human institutions.

The implementation of these technologies is not without challenge, requiring significant investments in computational infrastructure and a willingness to navigate complex structural trade-offs. Yet, the cost of inaction—measured in market volatility, institutional mistrust, and

societal inequality—is far higher. As we move toward an era of increasingly autonomous governance, the principles of accountability by design and cryptographic transparency must become the standard for all high-stakes AI systems. This research has demonstrated that a high-integrity architecture is both technically feasible and socially necessary, providing a blueprint for a future where financial AI is as responsible as it is efficient. By anchoring our technological progress in the principles of verifiability and provenance, we can build a financial infrastructure that is truly resilient, fair, and aligned with the long-term interests of society.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.

2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

3. Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. Machine Learning, 81(2), 121-148.

4. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

5. Boyd, D., & Crawford, K. (2012). Critical questions for big data: Interrogations of a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662-679.

6. Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. Big Data & Society, 3(1).

7. Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133).

8. Costan, V., & Devadas, S. (2016). Intel SGX explained. Cryptology ePrint Archive.

9. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

10. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

11. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12),

86-92.

12. Gulzar, M. A., Interlandi, M., Yoo, S., Tetali, S. D., Tyson, B., Kim, M., & Millstein, T. (2016). BigDebug: Debugging primitives for interactive big data processing in Spark. Proceedings of the 38th International Conference on Software Engineering.

13. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

14. Hendrickson, J. M. (2023). Algorithmic trading and systemic risk: A review of the infrastructure. Journal of Financial Regulation and Compliance, 31(4), 450-468.

15. Juels, A., Kosba, A., & Shi, E. (2016). The ring of Gyges: Investigating the future of criminal smart contracts. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.

16. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 1633-1705.

17. Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2), 1-16.

18. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

19. Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton University Press.

20. O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

21. Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

22. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., ... & Chua, T. S. (2025). Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. arXiv preprint arXiv:2511.17282.

23. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Jost, J., & Denton, E. (2020). Saving Face: Investigating the ethical concerns of facial recognition auditing. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

24. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

25. Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-based access control models. Computer, 29(2), 38-47.

26. Seligman, J. S., & Lewis, J. (2022). Governance of autonomous financial systems. Review of Financial Studies, 35(8), 3900-3945.

27. Stallings, W. (2017). Cryptography and Network Security: Principles and Practice. Pearson.

28. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

29. Taddeo, M., & Floridi, L. (2018). Regulating algorithms: Trust, transparency and accountability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133).

30. Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. arXiv preprint arXiv:2007.05558.

31. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841.

32. Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., ... & Zumar, R. (2018). Accelerating the machine learning lifecycle with MLflow. IEEE Data Engineering Bulletin, 41(4), 39-45.