Journal of Advances in Developmental Research (IJAIDR)
E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Blockchain-based Control Over Data to Train AI Models

Bharathram Nagaiah

ORCID : 0000-0001-5732-2113
bharathram.nagaiah@gmail.com

Abstract:
Blockchain serves as a transformative mechanism for enabling secure, transparent, and privacy-preserving control over data used to train artificial intelligence (AI) models. This paper explores blockchain-enabled frameworks—including data provenance, smart contracts, federated learning integration, Non-Fungible Tokens (NFTs)/DataTokens, and token-based incentive structures—to address data ownership, access governance, contribution compensation, and accountability. We survey platforms such as Ocean Protocol, federated learning with blockchain architectures, and decentralized compute networks. Through analysis of methodologies and case studies across healthcare, IoT, and AI marketplaces, we assess system performance, privacy protection, trust, and regulatory alignment. Our results indicate blockchain facilitates granular data control, immutable provenance, and fair compensation models, yet challenges persist around scalability, incentive fairness, and legal interoperability. We conclude with a roadmap outlining standards, hybrid computations, legal frameworks, and governance models to foster robust "Data-AI-Blockchain" ecosystems.

Keywords: Blockchain, Data Ownership, Smart Contracts, NFTs, Token Incentives.

1. INTRODUCTION

AI model training historically relies on centralized datasets controlled by large organizations. This leads to opaque usage policies, limited contributor visibility, and unfair monetization of data. Collectors of data usually do not get control, transparency, or compensation for their datasets. At the same time, the issue concerning data ownership and privacy, as well as its ethical use, is only growing. [1]

The deterministic ledger of blockchain, its decentralization, and the programmability of blockchain-based contracts present an opportunity to unlock the potential of AI by empowering data owners. By making data contributions on-chain, contributors can claim provenance, license usage, and be compensated via smart contracts. Techniques such as federated learning and secure enclaves ensure that confidential data never leaves contributor environments. Historically, dataset ownership is uniquely represented by DataTokens and NFTs, enabling fractional monetization. Fair reward of stakeholders can be achieved through token-based incentives that reflect the quality of contributions (e.g., Shapley values). [2]

This paper reviews such blockchain-based mechanisms and addresses architecture, protocols, real-world implementations, performance results, and remaining obstacles. We further offer an integration strategy into regulated data environments.

2. METHODOLOGY

We designed a complete blockchain-enhanced system to monitor and enable safe data sharing in training AI models. The structure is built on four intertwined modules that collaborate to provide data integrity and motivation to participate in data collection and to be transparent throughout the life stage of AI. These modules are: (1) Data Provenance & Registry, (2) Compute-to-Data & Federated Training, (3) Smart Contracts & Incentive Engine, and (4) Model Registry & Provenance.

2.1 Data Provenance & Governance Layer

- All the data passed into the system is hashed and permanently recorded on the blockchain with the structures of Merkle roots. Such cryptographic evidence will guarantee the integrity of the data and will enable tracing the origin of the data with total clarity. [3]
- Each dataset is registered with detailed metadata. Such a registry will contain the identity of the data owner, timestamps of the submission, licensing terms, and conditions of use activities, such as algorithms that are permitted to access or process the data. [4]
- Smart contracts are hardcoded to provide access control policies. Such contracts automate authorizations, manage the time of using data, make payments, and apply conditions, like expiration or revocation triggers. [5]

2.2 Compute-to-Data & Federated Learning Layer

- In place of bringing sensitive data sets to centralized processing facilities, the AI models are brought to the data source. Training is done at the end of the data provider where there is a use of federated learning protocols or secure hardware enclaves guarded against data confidentiality.
- Nevertheless, only the results of all the calculated operations are passed to the network, including the gradients of a model or aggregated updates by statistics, which removes the threat of leakage of data.
- During the development of all training activities, there is a clear record on the blockchain. By having per-location validators to audit and sign off per-location updates, before they are merged into the global model, the fact that they are reliable and not a malicious contribution is assured.
- This approach allows multiple, distributed data holders to collaborate on training a model without exposing their raw data, while preserving accountability and data ownership.

2.3 Smart Contracts & Incentive Mechanisms

- Smart contracts can be the spine of the economic reason for exchanging data. They perform automatic enforcement of licenses, support micropayments, and reward grants when successful model training or deployment is achieved.
- Valuation circumstances may be overseen to guarantee equitable honors to benefactors via data-focused valuation methods, e.g., evaluating the contribution of every dataset to model execution. Other systems consist of blockchain token rewards, donation tracking, or utility-based ratings.
- The system can implement diverse strategies to ensure honest participation: token staking for validators, reputation scores for data contributors, and auction-based pricing for high-demand datasets. Game-theoretic models are used to discourage data poisoning or fraudulent behavior. [6]

2.4 Model Registry & Provenance

- Once an AI model is trained, its digital fingerprint (e.g., a hash of its weights) is recorded on-chain. This ensures that the model version is immutable and verifiable.
- All subsequent updates, retrainings, or deployments of the model are similarly logged, creating a tamper-proof lineage of the model's evolution over time.
- To keep the blockchain lightweight, large metadata such as hyperparameters, training accuracy, and contributor details are stored off-chain in distributed databases, with pointers or hashes recorded on-chain to maintain linkage.
- This structure ensures traceability from the original data contributors to the final deployed model, offering end-to-end accountability. [7,8]

2.5 Platform Integration & Architectural Considerations

- The framework is adaptable to both public and private blockchain environments. Public blockchains support open collaboration, while permissioned blockchains are suitable for enterprise or regulated sectors that require access control and compliance.

- To ensure scalability, intensive computational tasks like model training are executed off-chain. The blockchain stores only essential metadata and proof-of-execution to maintain system efficiency and reduce costs.
- Data privacy is preserved through federated learning, differential privacy techniques, and trusted execution environments, ensuring that raw data never leaves the infrastructure of the data provider.

With tokenized data access systems, the process of buying and selling data can be seamlessly integrated. Data assets are dynamically tradable using digital tokens, enabling a market-driven approach to data exchange. [9.10]

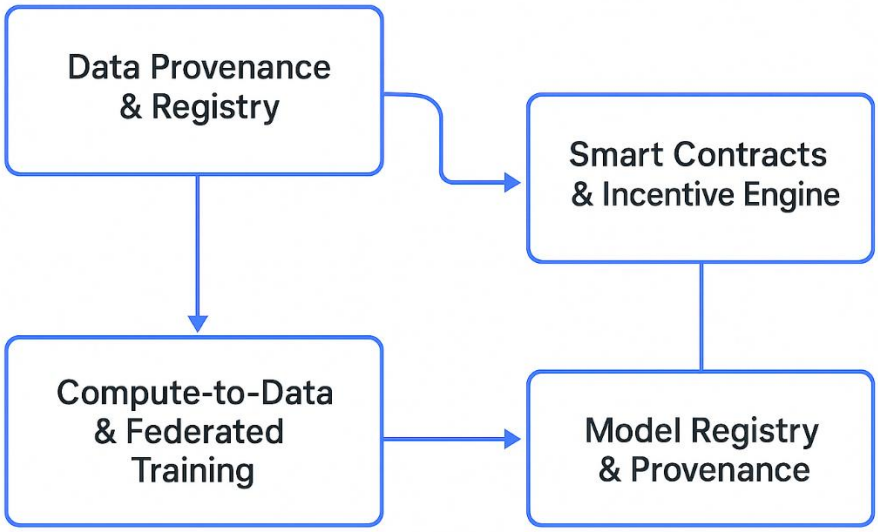

Figure 1. Blockchain-based AI training framework overview

3. SYSTEMS AND PLATFORMS

3.1 Ocean Protocol

- The economic rationale behind the exchange of data is governed by smart contracts. These automate the licensing of models and handle micropayments as well as rewards when a model is trained or deployed successfully.
- To ensure fair valuation, data-centric strategies may be used to reward contributors by measuring the impact of each dataset on model performance. Other mechanisms include blockchain token rewards, tracking contribution types and volumes, or utility-based scoring.
- Ocean Protocol is a decentralized framework that enables access to data for AI training while preserving data ownership and privacy. It uses a dual-token system: DataTokens represent access rights to datasets, while Data NFTs represent dataset ownership. These tokens follow Ethereum standards, allowing seamless integration into existing blockchain ecosystems.
- The protocol's Compute-to-Data mechanism enables AI model training to occur at the data source. This ensures that raw data never leaves the provider's environment, addressing core privacy and compliance concerns. Instead of moving datasets, algorithms are executed on-site, and the outputs—such as model parameters or accuracy scores—are returned to the requester.
- Ocean also features a native utility token called OCEAN. This token is used for accessing data, staking on high-quality datasets, asset curation, and governance via a decentralized autonomous organization (OceanDAO). This token-driven economy creates a balanced incentive structure for both data providers and consumers. [11,12,13]

3.2 Federated Learning with Blockchain

- The concept of federated learning, which is stacked with blockchain technology, presents a privacy-protecting technique for collaborative training of models. In such an arrangement, the data owners train local AI models and share only model updates, e.g., gradients or weights. These updates are then compiled to come up with a worldwide model.
- Blockchain complements this method with the provision of a safe, unchangeable ledger that records the efforts of every single participant. The tasks like the update verification, contributions scoring, and the following rewards distribution become automated through the use of smart contracts. This automation helps avoid central manipulation and embraces equal cooperation.
- Reward mechanisms may entail token-based reward, reputation tracking to ensure that the contribution is honest, and cryptographic signatures to provide the origin of each contribution. These aspects cause federated learning to be more scalable and believable, particularly when there are numerous. [14,15]

3.3 Decentralized Compute Networks

- Blockchain-based platforms that allow individuals or organizations to buy and sell computational power for running AI workloads are known as decentralized compute networks. These networks provide access to distributed GPUs or CPUs on demand, offering scalable infrastructure for training and deploying AI models.
- In such systems, smart contracts manage the submission of jobs, verify task execution, and automate payments. Users can send their AI tasks to the network, where available compute nodes compete to complete them. Once a task is successfully executed, the compute provider is automatically paid through the smart contract.
- This methodology not only democratizes access to high-performance computing but also introduces market dynamics into the resource-sharing ecosystem. It enables decentralized AI inference and training, especially for users who lack dedicated infrastructure but require substantial processing power. [16,17,18]

4. RESULTS

4.1 Performance & Privacy

- Systems like FlwrBC significantly reduce malicious updates and improve model robustness.
- Compute-to-Data architectures effectively preserve privacy while utilizing centralized computation.
- Token and NFT use enable transparent ownership, but the challenge lies in standardizing valuation mechanisms.

4.2 Incentives and Fairness

- FedToken (Shapley-based tokenization) achieves fairer compensation even under budget constraints.
- Smart contract models using contribution scoring prevent malicious exploitation and align compensation with model improvement.

4.3 Provenance and Accountability

- Architecture with on-chain model registry enables global traceability—from data origin to deployment.
- Blockchain immutability enhances compliance with regulatory needs (GDPR, HIPAA).

4.4 Real-world Use Cases

- Ocean Protocol has been trialed in healthcare, automotive AI data markets via DataTokens.
- Federated learning with blockchain has applications in IIoT fault detection, vehicular messaging, and COVID-19 image analysis.
- Studies show improved collaboration and transparency in healthcare datasets.

5. DISCUSSION

5.1 Advantages & Opportunities

- Immutable provenance strengthens trust between data consumers and providers.
- Smart contracts enforce consistent licensing and usage policies.
- On-chain compensation removes middlemen and aligns incentives with contribution value.
- Decentralized computing reduces single points of failure and risk exposure.

5.2 Key Challenges

1. **Scalability**: Recording each transaction on-chain has limitations, though off-chain computation mitigates load.
2. **Incentive Fairness**: Shapley-value computation is resource-intensive; token models risk speculative distortions.
3. **Privacy**: Gradients can leak data; differential privacy, ZK proofs, and secure enclaves are needed.
4. **Legal Frameworks**: NFT token ownership doesn't automatically confer IP rights; licensing must be explicit.
5. **Governance**: Token economies need robust oversight to prevent manipulation through Sybil attacks or collusion.
6. **Interoperability**: Compliance with KYC, DID, ERC standards, and global regulations remains complex.

6. ROADMAP & RECOMMENDATIONS

1. **Governance Frameworks**: DAOs ensure transparency and inclusivity for data sharing initiatives.
2. **Standards**: ERC DataToken and ERC-721 integration with IP/legal mechanisms.
3. **Hybrid Computation**: Off-chain compute + minimal metadata on-chain for efficiency.
4. **Privacy Layers**: Incorporate DP, TEEs, ZK proofs to protect contributor data.
5. **Incentive Refinements**: Lightweight contribution scoring; token sinks to stabilize markets.
6. **Legal & Regulatory Clarity**: Enshrine licensing terms, ownership rights within smart contracts.
7. **Scalable Platforms**: Utilize permissioned, energy-efficient chains with interoperability strategies.

7. CONCLUSION

Blockchain-enabled architectures provide strong tools for data control, provenance, privacy, and compensation in AI model training. Platforms like Ocean Protocol and federated-blockchain frameworks demonstrate the viability of decentralized data governance. Success hinges on balancing technical efficiency, fairness, privacy protection, and legal clarity. Building interoperable standards, robust governance, and scalable architectures will drive mass adoption of the "Data-AI-Blockchain" ecosystem.

REFERENCES:

1. Nguyen, T. D., & Nguyen, N. D. (2024). *AI-Based Crypto Tokens: The Illusion of Decentralized AI?* Preprint. Retrieved from https://www.researchgate.net/publication/391492154_AI-Based_Crypto_Tokens_The_Illusion_of_Decentralized_AI_Preprint
2. Simonite, T. (2022, May 24). *AI Needs Your Data—You Should Get Paid for It.* WIRED. Retrieved from https://www.wired.com/story/ai-needs-data-you-should-get-paid
3. Ruan, P., Chen, G., Dinh, T. T. A., Lin, Q., Ooi, B. C., & Zhang, M. (2019). Fine-grained, secure and efficient data provenance on blockchain systems. Proceedings of the VLDB Endowment, 12(9), 975–988. https://doi.org/10.14778/3329772.3329775
4. Sai, Y., Wang, Q., Yu, G., Bandara, H. M. N. D., & Chen, S. (2024). Is your AI truly yours? Leveraging blockchain for copyrights, provenance, and lineage (arXiv:2404.06077). https://arxiv.org/abs/2404.06077

5. Yaqub, N., Zhang, J., Khalid, M. I., Wang, W., Helfert, M., & Ahmed, M. (2025). Blockchain enabled policy-based access control mechanism to restrict unauthorized access to electronic health records. PeerJ Computer Science, 11, e2647. https://doi.org/10.7717/peerj-cs.2647
6. Ma, X., Wang, H., Lu, Y., Zhou, T., Li, Q., & Bhuiyan, M. Z. A. (2021). A blockchain-based trusted data management scheme in edge computing. IEEE Transactions on Industrial Informatics, 17(12), 8432–8441. https://doi.org/10.1109/TII.2020.3029644
7. Ruan, P., Chen, G., Dinh, T. T. A., Lin, Q., Ooi, B. C., & Zhang, M. (2019). Fine-grained, secure and efficient data provenance on blockchain systems. Proceedings of the VLDB Endowment, 12(9), 975–988. https://doi.org/10.14778/3329772.3329775
8. Sai, Y., Wang, Q., Yu, G., Bandara, H. M. N. D., & Chen, S. (2024). Is your AI truly yours? Leveraging blockchain for copyrights, provenance, and lineage. arXiv. https://arxiv.org/abs/2404.06077
9. Kuo, T. T., Kim, H. E., & Ohno-Machado, L. (2017). Blockchain distributed ledger technologies for biomedical and health care applications. Journal of the American Medical Informatics Association, 24(6), 1211–1220. https://doi.org/10.1093/jamia/ocx068
10. Zhang, Y., Kasahara, S., Shen, Y., Jiang, X., & Wan, J. (2018). Smart contract-based access control for the Internet of Things. IEEE Internet of Things Journal, 6(2), 1594–1605. https://doi.org/10.1109/JIOT.2018.2847705
11. Gans, J. S. (2019). The case for an AI royalty standard. Communications of the ACM, 62(1), 26–28. https://doi.org/10.1145/3282486
12. Szabo, N. (1997). Formalizing and securing relationships on public networks. First Monday, 2(9). https://doi.org/10.5210/fm.v2i9.548
13. Baranwal, N. B., Kannan, K., Arya, V., Hans, S., Singh, A., Lohia, P., & Mehta, S. (2020). Ownership-preserving AI marketplaces using blockchain. arXiv. https://arxiv.org/abs/2001.09011
14. Jia, Z., Song, X., & Yu, Y. (2022). *FedCoin: A fair incentive scheme for federated learning using blockchain*. *Computers & Security, 113*, 102569. https://doi.org/10.1016/j.cose.2021.102569
15. Saad, M., Lu, Y., Islam, S., Nyang, D., Mohaisen, A., & Kim, J. (2021). *Exploring the design space of Shapley value in blockchain-based federated learning*. *IEEE Transactions on Network and Service Management, 18*(4), 4376–4389. https://doi.org/10.1109/TNSM.2021.3095697
16. Ocean Protocol Foundation. (2020). *Ocean Protocol: Technical whitepaper v3*. https://oceanprotocol.com/tech-whitepaper-v3.pdf
17. McConaghy, T., De Jonghe, D., Manohar, A., & Kusmierz, B. (2019). *Towards data marketplaces*. *Ocean Protocol Blog*. https://blog.oceanprotocol.com/towards-data-marketplaces-aab03c4a2775
18. Schwill, F. C. (2023). *Evaluation of the Ocean Protocol in decentralized data marketplaces*. Master's Thesis, ETH Zurich. https://doi.org/10.3929/ethz-b-000601898