CSK JOURNAL OF HUMANITIES AND SOCIAL SCIENCES
Research Article
Volume 1| issue 1

Research on Intelligent Traceability Framework for Trusted Artificial Intelligence

Shiyu Sun[1] & Qinmi Sun[2*]

[1]*Zhengzhou University of Light Industry, Zhengzhou*
[2]*Henan Institute of Metrology, Zhengzhou, China*

[*]**Corresponding Author:** Qinmi Sun, Henan Institute of Metrology, Zhengzhou, China.

Submitted: 16 November 2025 **Accepted:** 21 November 2025 **Published:** 27 November 2025

Citation: *Sun, S., & Sun, Q. [2025]. Research on Intelligent Traceability Framework for Trusted Artificial Intelligence. CSK J. Humanit. Soc. Sci. 1(1),*

Abstract
With the deep integration of artificial intelligence and big data technology, the self-learning ability of the system brings efficiency improvement, but problems such as data pollution, algorithm black box, and model drift exacerbate the difficulty of tracing. This article proposes a three-layer traceability framework (TVB-Trace) that integrates blockchain metadata anchoring, dynamic verification mechanism, and trusted execution environment. By constructing a verifiable data lineage graph and algorithm decision chain throughout the entire lifecycle, it achieves transparent supervision of AI self-learning systems. Experiments have shown that this framework can improve data traceability accuracy to 99.2% and enhance model decision interpretability by over 40%. (Keywords: artificial intelligence traceability, blockchain, trusted computing, self-learning system).

Introduction

Tracing back is a commonly used verb in Chinese, which originally refers to searching for the source of a river upstream, and is extended to explore the essence or origin of things. The term can be traced back to the field of geographical exploration and has since been widely used in academic research, cultural research, and other contexts. The semantic core emphasizes the exploration process upstream, which includes both the pursuit of spatial dimensions and the historical retrospective of temporal dimensions. This article refers to the real guarantee of trustworthy data sources for artificial intelligence.

The self-learning capability of artificial intelligence (AI)systems, while enhancing efficiency, has precipitated a severe trust crisis due to data pollution, algorithmic black - boxing, and model drift. In recent years, AI has been widely applied in various fields such as finance, healthcare, and transportation. For example, in the financial sector, AI - based credit scoring models are used to assess the creditworthiness of borrowers. However, the self-learning nature of these models can lead to unexpected results [1- 3].

Current research primarily focuses on single-stage traceability (e.g.,data source verification or model interpretability), lacking a comprehensive lifecycle governance framework. Technical bottlenecks in traceability exist, as most existing methods address only one aspect of the problem—such as validating data sources—while neglecting changes in models during the learning process.

Three main drawbacks of current AI systems are:

Data Pollution

Dynamic data streams in self-learning systems make it challenging for traditional databases to track adversarial operations (e.g., injecting malicious samples). For instance, in social media sentiment analysis models, attackers can inject fake positive or negative comments to manipulate outputs. Traditional databases lack real-time detection capabilities for such dynamic, adversarial data.

Model Drift

Iterative algorithm updates often deviate from initial objectives, obscuring decision logic. For example, a fraud detection model designed to identify common fraud patterns may shift focus to new data patterns, reducing detection rates for original fraud types.

Black Box Dilemma

The non-linear characteristics of deep neural networks hinder transparent reconstruction of decision chains. For instance, in convolutional neural network (CNN)-based image recognition, it is difficult to understand how the model identifies specific objects.

TVB-Trace constructs a verifiable lineage graph from data ingestion to model output, addressing the "trust gap" in autonomous AI systems.

Tvb-Trace Framework Design

Architectural Overview

To address these challenges, the TVB-TRACE framework is designed with three layers:

Metadata Anchoring Layer: Blockchain technology records data sources, preprocessing operations, and model versions, generating globally unique data fingerprints (Merkle tree structure)[4]. The Merkle tree enables efficient integrity verification of large datasets. For example, in distributed storage systems, it quickly identifies modified data blocks.

Dynamic Validation Layer: Lightweight validation nodes monitor data streams and model states in real time, triggering alerts and version rollbacks. These nodes can be deployed across distributed AI systems to detect anomalies like sudden data distribution shifts or performance declines.

Trusted Execution Layer: Core algorithms run in Trusted Execution Environments (TEEs), ensuring auditable training and inference processes. For example, TEEs like Intel SGX protect model training from external interference, enabling transparent audits.

Key Technical Implementations

Provenance Graph Construction: Directed Acyclic Graphs (DAGs) capture cross-chain relationships between multimodal data (e.g., text, images, videos). DAGs visualize complex data interactions, such as associations between multimedia elements in processing systems.

Smart Contract-Driven Validation: On-chain rules (e.g., data cleaning thresholds, performance degradation metrics) automate validation workflows. For instance, smart contracts can trigger automatic data cleanup if quality falls below predefined standards.

Interpretability Enhancement: SHAP values combined with blockchain logs visualize decision paths and feature contributions, improving model transparency.

Experiments And Evaluation

Experimental Setup

Datasets:Lending Club Financial Dataset: Contains loan applications, borrower credit records, and repayment histories. Includes 5% adversarial samples simulating malicious loan manipulation.

MIMIC-III Medical Dataset: Includes patient clinical data (e.g.,vital signs, lab results). Contains 5% adversarial samples mimicking data errors or tampering.

Baselines: Compared with centralized logging systems and standalone blockchain solutions..

Results

Traceability Accuracy:TVB Trace achieved 99.2% adversarial sample detection, outperforming baselines by 23% (Table 1).

Table 1: Comparison of Traceability Accuracy

Framework	Detection Rate for Adversarial Samples	Improvement over Baselines
TVB - Trace	99.20%	23%
Centralized Logging System	76.20%	-
Standalone Blockchain Solution	73.10%	-

Trust Enhancement:The decision chain visualization feature provided by TVB's Trace framework has increased users' trust in the model output by 42%. In artificial intelligence systems, especially in key application areas such as finance and healthcare,User trust is crucial.

In our framework, the combination of SHAP values and blockchain logs allows us to visually display the model's decision path and its feature contributions. This transparency helps users understand how the model makes decisions and which factors influence those decisions. For example, in the loan approval model, users can clearly see which features, such as credit score or income, have the greatest impact on the approval results. Therefore, visualizing the decision chain is key to establishing user trust in the AI system.

In addition, blockchain based metadata anchoring ensures the authenticity of the data used in the model, which is a major concern for users when using AI systems. Blockchain technology can effectively address this issue.Table 2 shows a comparison of user trust in the model output before and after using the TVB Trace framework.

Table 2: Comparison of Trust Enhancement

Scenario	User Trust in Model Outputs
Before using TVB - Trace	38.00%
After using TVB - Trace	80.00%

Performance Overhead

In the TVB-Trace framework, TEE operations initially resulted in 18% training latency. Although TEE provides a secure environment for algorithm execution, additional security measures also incur certain costs.

However, through parallel computing, we successfully reduced the training latency to 7%. Parallel computing refers to breaking down training tasks into multiple subtasks and running these subtasks simultaneously on multiple processors or cores. By utilizing parallel computing technology, we can fully utilize existing computing resources to accelerate training speed. Table 3 shows the comparison of training latency before and after using parallel computing.

Table 3: Comparison of Performance Overhead.

Scenario	Training Latency
Without parallel computing	18.00%
With parallel computing	7.00%

effectiveness of the TVB - Trace framework in terms of traceability accuracy, trust enhancement, and performance. The combination of blockchain metadata anchoring, dynamic validation mechanisms, and trusted execution environments provides a comprehensive solution for AI traceability, addressing the challenges posed by data pollution, algorithm black - boxing, and model drift

.

Conclusion

Due to the limitations of technological progress, there are still technical deficiencies, one of which is cross chain interoperability: the standardization of metadata on heterogeneous blockchains has not yet been resolved. The second is TEE hardware dependency: edge device compatibility requires software and hardware co design.

However, the TVB Trace framework constructs a lifecycle aware governance model for self-learning AI through three layers of traceability. As long as the technological deficiencies are gradually addressed, future development directions including cross institutional data collaboration in federated learning and blockchain hybrid technology will have broad prospects.

References

1. Yang, Z., Zhang, R., & Zhang, L. (2025). Uncovering the black box of medical image analysis algorithms: The latest advances in explainable artificial intelligence in medical image analysis. Chinese Science Bulletin, 1.
2. Huang, X., Zhou, X., & Yang, P. (2012). Research on key technologies of a new decision system based on human service. Computer Applications and Software, 2, 19–21.
3. He, Y., & Li, R. (2022). International research hotspots and trends of adaptive learning platforms from 2010 to 2021: Based on CiteSpace visualization analysis. Journal of Yunnan Normal University (Natural Sciences Edition), 2, 67–74.
4. Xie, P., Ran, Y., & Feng, T. (2025). Industrial OT network access authorization traceability method based on balanced Merkle tree. Journal on Communications, 4, 282–294.