Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 19.07.2025 Revised: 27.10.2025 Accepted: 23.12.2025

UDC 004.8:[004.7:34] DOI: 10.31649/vitce/3.2025.173

The EBAT architecture: An explainable blockchain for legal AI audits

Oleksii Shamov*

Intelligent Systems Researcher
Human Rights Educational Guild
18010, 40/28 Rizdviana Str., Cherkasy, Ukraine
http://orcid.org/0009-0009-5001-0526

Abstract. The integration of artificial intelligence into high-stakes domains like the justice system presents the "black box problem", where algorithmic opacity undermines fundamental legal principles and current blockchain-based auditing solutions fail to bridge the critical gap between a record's technical integrity and its value as interpretable legal evidence. This research aimed to develop and theoretically substantiate a novel audit system architecture that synergistically combines the cryptographic reliability of blockchain with the interpretive power of Explainable Artificial Intelligence (AI) to produce logs of AI decisions that are not only immutable but also legally significant and human-understandable. The methodology involved a systematic analysis and synthesis, including a review of publications from scientometric databases, an analysis of legal standards for digital evidence, and conceptual architectural design methods for information systems. The study proposed a new hybrid architecture, the "Explainable Blockchain Audit Trail", specifically designed to resolve this challenge. Its core novelty lies in a three-tiered structure that first mandates the generation of human-readable counterfactual explanations for every AI decision. Second, a complete and self-sufficient evidence package - containing the input data, model specifications, and the generated explanation - is securely stored in decentralised off-chain storage to ensure its integrity and availability. The third tier then creates an immutable "trust anchor" for this package on a permissioned blockchain, cryptographically linking all components and providing a permanent, tamper-proof record of the event. This comprehensive model ensures complete reproducibility of the decision-making process and establishes a robust, objective basis for judicial review and appeal. The proposed architecture provides a crucial theoretical foundation for developing practical tools for judges, lawyers, and regulators, ultimately aiming to enhance transparency and protect citizens' rights in the age of algorithmic decision-making by offering concrete mechanisms to challenge opaque conclusions

Keywords: legal tech; audit trail; digital evidence; immutable logs; justice

Introduction

The modern world is experiencing the fourth industrial revolution, driven by artificial intelligence (AI). AI technologies are penetrating the most conservative and responsible spheres of human activity, including finance, healthcare, and, most importantly, the justice system. The potential for optimising legal processes, analysing large arrays of case law, and assisting in decision-making is substantial. However, this technological expansion carries a systemic risk known as the "black box problem", where the internal logic of complex machine learning models remains hidden from human analysis. This opacity is not just a technical flaw but an existential threat to the rule of law, as it undermines principles of a fair trial and the right to a reasoned explanation. As highlighted in a review by S. Verma (2019) on the societal impact of algorithmic systems, such opaque models can perpetuate and even amplify existing biases, leading to discriminatory outcomes and deepening inequality, creating what are effectively "weapons of math destruction". Consequently, achieving "Trustworthy AI" is impossible without deeply integrated explainability mechanisms, a conclusion strongly supported by the comprehensive review from S. Ali *et al*. (2023), who argued that trust is unattainable until we can adequately understand and scrutinise AI-driven conclusions.

In response to these challenges, researchers have explored the use of blockchain technology to create immutable,

Suggested Citation:
Shamov, O. (2025). The EBAT architecture: An explainable blockchain for legal AI audits. *Information Technologies and Computer Engineering*, 22(3), 173-181. doi: 10.31649/vitce/3.2025.173

*Corresponding author

cryptographically protected audit trails for AI systems. A foundational review by K. Salah *et al*. (2019) outlined the significant potential of blockchain to enhance AI by providing a decentralised and tamper-proof mechanism for recording data and model actions, thereby addressing key issues of accountability and data integrity. The legal community is also actively examining these developments. For instance, X. Wang *et al*. (2024) explored the procedural implications of blockchain-based evidence in U.S. judicial processes, concluding that while blockchain's cryptographic security offers strong proof of integrity, its admissibility in court is still complex and requires clear standards for validation. This legal challenge is further emphasised by R. Bharati *et al*. (2024), whose work on digital evidence highlighted the significant hurdles that remain in ensuring its admissibility, given the ease of manipulation and the difficulty of establishing an unbroken chain of custody for digital artifacts. These issues are not merely academic; they are at the forefront of regulation, as analysed by S. Ramos & J. Ellul (2024), who detail how distributed ledger technology can be a critical tool for organisations seeking to comply with the stringent transparency and record-keeping requirements of the EU AI Act.

Despite these advancements, a critical aspect often remains underexplored: the very nature of what constitutes a valid explanation. The work of G. Vilone & L. Longo (2021) delved into the core notions of explainability, arguing that it is not a monolithic concept. They analysed various approaches to evaluating explanations, concluding that an effective explanation must be tailored to the context and the user, moving beyond purely technical metrics to ensure genuine human comprehension and utility.

A review of the current literature revealed a significant research gap. While substantial work has been done on using blockchain to ensure the technical integrity of AI logs, and a separate body of research has established the critical need for human-centric explainability, there is a lack of integrated frameworks that combine both elements to create legally admissible, interpretable evidence. Current models often focus on proving that a log has not been tampered with, but they do not address the equally important question of whether the decision recorded in that log is understandable and justifiable. This unexplored intersection between cryptographic immutability and semantic interpretability creates the urgent need for the present study.

Therefore, the purpose of this article was to develop and theoretically substantiate a new hybrid architecture for an AI decision audit system that synergistically integrates blockchain technology and Explainable AI (XAI) methods to create complete, immutable, and human-understandable evidence. To achieve this purpose, this article set out three primary objectives: first, to design a conceptual model of a hybrid architecture, the "Explainable Blockchain Audit Trail" (EBAT), which combines a permissioned blockchain, decentralised storage, and a mandatory explanation generation module; second, to analyse the potential of specific XAI methods, particularly counterfactual explanations, as the core tool for transforming technical logs into legally significant evidence within this architecture; and finally, to justify how the proposed architecture solves the "black box" problem in a legal context by meeting key requirements for digital evidence, including integrity, authenticity, and interpretability. The scientific novelty of this work lies in the development of the original three-tiered EBAT architecture, which synergistically combines three technologies (XAI, IPFS (InterPlanetary File System), Hyperledger Fabric) into a single system designed to meet the demands of the justice system by shifting the focus from simply recording a decision to recording its justification.

Literature Review

The problem of trust and accountability in artificial intelligence systems is multidisciplinary and is addressed in the works of many scholars in engineering, law, and social sciences. A review of the literature allowed for the identification of several key areas relevant to the topic of this study: the technical implementation of blockchain for data integrity, the legal standards for digital evidence, and the theoretical foundations of XAI. The first area concerns the technical aspects of using blockchain technology to ensure data integrity. One of the pioneering works in this domain was presented by A. Sutton & R. Samavi (2018), who proposed a conceptual framework for auditing AI systems using a tamper-proof log on a public blockchain. Their research established the foundational principle that cryptographic hashing of AI decisions and their inputs could provide a verifiable trail, though they also acknowledged the significant scalability and cost limitations inherent in using public ledgers like Bitcoin for high-frequency operations. Building on these early concepts, the review by K. Salah *et al*. (2019) provided a comprehensive map of the synergy between AI and blockchain. The authors systematically identified the main challenges and opportunities, concluding that while blockchain offers transformative potential for creating transparent and auditable AI systems, critical issues such as scalability, data privacy, and interoperability must be addressed through sophisticated architectural design. This idea was extended in broader research on blockchain-based trust management, as detailed in the survey by Y. Liu *et al*. (2023), which delved into the specific application of blockchain for trust management in the Internet of Things (IoT). Their work is highly relevant as it analysed how blockchain's decentralised and tamper-proof nature can establish verifiable trust between autonomous devices without a central authority, a principle directly applicable to creating trust in autonomous AI decisions in a legal setting. More recent works focused on specific regulatory contexts. For example, S. Ramos & J. Ellul (2024) directly connect distributed ledger technology to the stringent requirements of the European Union's AI Act. They argued that for high-risk AI systems, blockchain is not merely a technical option but a vital tool for demonstrating compliance, providing an

immutable record of the model's lifecycle, training data, and decision-making processes as required by the regulation. The practical application of these principles is now being actively implemented in various data-sensitive fields. M. Faruk *et al*. (2023) proposed a specific framework for securing electronic health records, utilising smart contracts and the InterPlanetary File System (IPFS) to ensure that patient data remains both confidential and integral, demonstrating a tangible use case for the hybrid on-chain/off-chain model. Similarly, Y. Zhang *et al*. (2023) explored the use of blockchain in federated learning, highlighting its role not only in preserving privacy but also in achieving verifiable fairness, a crucial concept for legal AI where algorithmic bias is a primary concern.

The second important area of research relates to legal standards and the admissibility of digital evidence. The work by X. Wang *et al*. (2024) provided a detailed analysis of how blockchain records might be treated within the U.S. judicial system. They concluded that while the cryptographic properties of blockchain provide a powerful technical argument for the authenticity and integrity of evidence, its legal admissibility is not automatic and hinges on procedural rules and the judiciary's evolving understanding of the technology. Their analysis underscored that a technical solution alone is insufficient without a clear legal framework for its acceptance in court. The principles for handling digital evidence, famously outlined by the UK's Association of Chief Police Officers (ACPO) and detailed in the work of E. Casey (2011), established the gold standard for digital forensics, requiring that no action should change original data and that a full audit trail of all processes must be maintained. This legal precedent set a high bar for any system claiming to produce court-admissible evidence. This point is further elaborated by R. Bharati *et al*. (2024), who examined the broader challenges of digital evidence in legal proceedings. They emphasised the inherent fragility of digital artifacts and the critical importance of maintaining a verifiable and unbroken "chain of custody", concluding that traditional methods are often inadequate and that new technological solutions are needed to meet longstanding evidentiary standards.

The third, and key for this work, area is Explainable AI, where the focus shifts from the integrity of the record to the intelligibility of its content. One of the foundational works that popularised model-agnostic explanations was the article by M. Ribeiro *et al*. (2016), which introduced the LIME (Local Interpretable Model-agnostic Explanations) technique. Their key insight was that one could explain any black-box model's prediction by approximating it with a simpler, interpretable model in the local vicinity of that prediction, a powerful concept that opened the door to a wide range of explanation methods. The scientific community has since conducted significant work systematising these approaches, as detailed in the comprehensive survey by R. Guidotti *et al*. (2018). They provided a robust taxonomy of XAI methods, classifying them into different categories and analysing the types of explanation problems they solve, which helps in selecting the appropriate technique for a given context. However, for the legal sphere, counterfactual explanations hold particular value. In their influential work, S. Wachter *et al*. (2018) argued that it is precisely counterfactual explanations ("the decision would have been different if…") that best meet the General Data Protection Regulation's (GDPR) requirements for a "right to explanation". They posited that such explanations are intuitively understandable and provide individuals with actionable recourse, a feature often missing from other explanation types. Building on this, the work of G. Vilone & L. Longo (2021) offered a crucial theoretical foundation by deconstructing the very concept of "explainability". They provided a detailed taxonomy of different types of explanations and, most importantly, argue that the evaluation of an explanation's quality cannot be purely technical. Instead, it must be human-centric, focusing on whether the explanation is genuinely useful, understandable, and satisfactory to the end-user, which strongly supports the need for legally-oriented explanation methods over generic technical outputs.

Despite significant progress in each of these distinct areas, their synergistic integration remains insufficiently explored. Most research combining AI and blockchain focuses on the technical mechanisms of immutability while overlooking the semantic content of what is being recorded. Conversely, works within XAI deeply analyse the nature of explanations but rarely propose robust, cryptographically secure frameworks for their storage and verification as legal evidence. It is this gap between proving a record is unchanged and proving it is understandable that the present study aimed to fill.

Materials and Methods

This study was theoretical and conceptual, and its methodology was based on a comprehensive approach that combines several analytical and design methods. The main stages of the work and the justification for the choice of methods were as follows: systematic literature review and analysis of the state of the problem; analysis of legal and technical standards; conceptual architectural design; synthesis and justification.

Stage 1. At the first stage, a systematic literature review was conducted to identify key concepts, existing solutions, and unresolved problems at the intersection of AI, blockchain, and jurisprudence. The search for sources was carried out in leading scientometric databases, Scopus and Web of Science. The following keywords and their combinations were used: "explainable AI", "blockchain audit trail", "AI accountability", "legal tech", "digital evidence", "immutable logs", 'Hyperledger Fabric governance", "counterfactual explanations". The inclusion criteria were: publications for 2018-2025, high citation index (for foundational works), relevance to the research topic, and presence of a DOI. The exclusion criteria were: purely marketing articles, non-peer-reviewed works, and overly narrow technical reports without analysis of the broader context. This method

allowed for the formation of a deep understanding of the current state of research, the identification of a scientific gap, and the justification of the work's relevance.

Stage 2. The second step was a detailed analysis of existing standards governing the handling of digital evidence. The key principles outlined in the ACPO good practice guide for digital evidence (2012) from the UK's Association of Chief Police Officers and the requirements for evidence authentication under the U.S. Federal Rules of Evidence (FRE 901) were analysed (The U.S. Congress, 2024). The purpose of this analysis was to define clear legal requirements that any AI audit system must meet for its results to be considered admissible evidence in court. This method allowed for the formulation of a set of criteria that became the basis for the architectural design.

Stage 3. This was the central stage of the research, where the method of conceptual modelling and architectural design of information systems was applied. Based on the requirements formulated in the previous stages, a new hybrid architecture, the Explainable Blockchain Audit Trail, was developed. The choice of specific technological components (Hyperledger Fabric, IPFS) was justified by their technical characteristics and suitability for corporate and legal sector tasks. Specifically, Hyperledger Fabric was chosen for its support of private transactions, high throughput, and modular architecture, which are advantages over public blockchains like Ethereum for this application (Androulaki *et al.*, 2018). IPFS was chosen for its content addressing mechanism, which guarantees the integrity of off-chain data (Benet, 2014). This method allowed for the creation of a detailed theoretical model of the system.

Stage 4. In the final stage, a synthesis of the obtained results was carried out. The proposed EBAT architecture was analysed for compliance with legal requirements and its ability to solve the identified "black box" problem. A comparative analysis of the EBAT architecture with existing approaches described in the literature was conducted to demonstrate its scientific novelty and potential advantages. The chosen methodology is valid for theoretical research as it allows for a systematic analysis of the problem, the formulation of a set of requirements, and, based on them, the development and justification of a new conceptual solution. The described sequence of steps is logical and can be reproduced by other researchers for verification or further development of the proposed ideas.

Results and Discussion

Based on the analysis and the applied methodology, a new conceptual architecture, the Explainable Blockchain Audit Trail, was developed. The EBAT architecture is a complex hybrid system that operates on three interconnected tiers. Each tier uses specific technologies to perform its part of the task, and together they create a single, cohesive, and legally significant audit trail (Fig. 1).

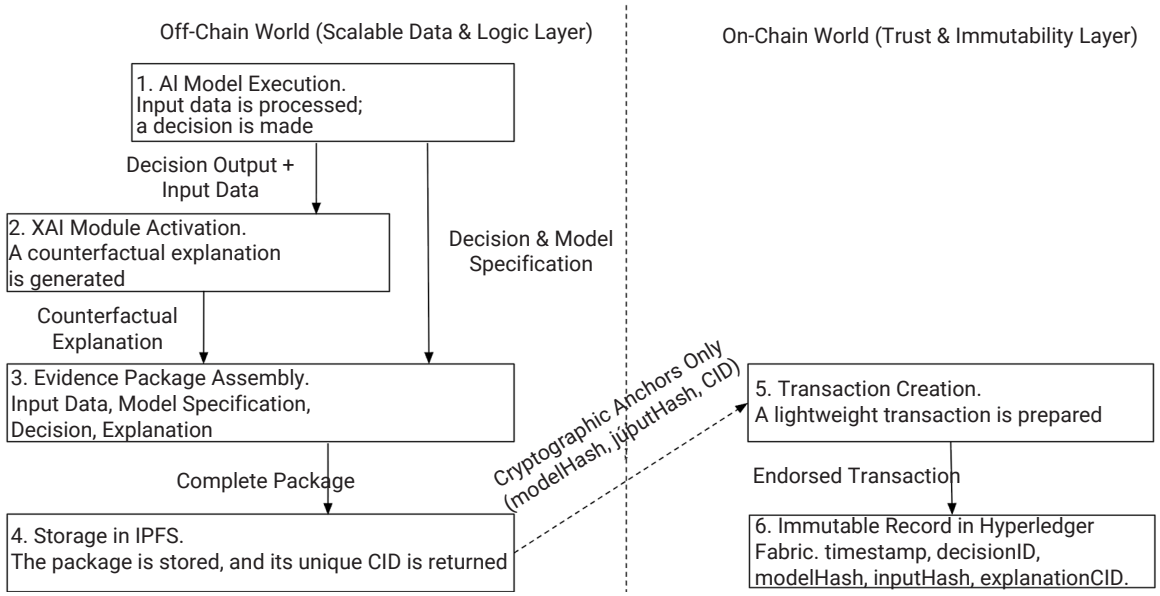

Figure 1. Overall architecture of the EBAT system

Source: completed by the author

Tier 1: Generation of the decision and the mandatory counterfactual explanation

This is the starting point and the ideological core of the entire architecture. The process begins when an AI system makes a decision with legal consequences (e.g., denying a loan, classifying evidence in a case, determining the risk of recidivism). In traditional systems, the logging process would end at this stage with the recording of technical parameters. In the EBAT architecture, however, this is just the beginning. Immediately after the decision is made, a mandatory integrated XAI module is activated This tier can be detailed into several sub-processes. First, the initial data (e.g., a loan applicant's form, digital evidence in a criminal case) passes through the main AI model. This model can

be anything from gradient boosting to a complex neural network but its conclusion (e.g., "deny", "approve", "high risk") is the trigger for the next step. Immediately after this conclusion is received, the same input data, along with the obtained result, is passed to the XAI module. The main requirement for this module is the generation of a counterfactual explanation. As justified in the work of S. Wachter *et al*. (2018), it is this type of explanation that is most valuable in the legal field, unlike other XAI methods such as SHAP (SHapley Additive exPlanations) or LIME.

While LIME (Ribeiro *et al*., 2016) explains which features were most important for a particular decision, a counterfactual explanation provides an actionable scenario. It answers the question: "What would need to be changed in the input data for the system to make an alternative, usually positive, decision?". This is a fundamental difference: instead of a passive statement ("you were denied because of low income"), the system provides an active instruction ("you would have been approved if your income was X higher"). For a lawyer, this means the ability to check whether the requirement for a "higher income" is discriminatory or unreasonable. For a citizen, it provides clear grounds for appeal.

Technically, generating counterfactual explanations is a complex optimisation task: the algorithm searches for minimal changes in the input feature vector that lead to a change in the model's classification while remaining plausible. Various algorithms exist for finding them, and the choice of a specific one depends on the type of AI model. It is important that the generation of such an explanation is not an option but a mandatory, atomic part of the decision-making process. Without the successful generation of an explanation, the entire process is considered incomplete and cannot be recorded. This ensures that no "silent" decisions without justification appear in the system.

Tier 2: Formation of the evidence package and its storage in decentralised storage

Since storing large volumes of data directly on the blockchain is technically inefficient and economically unfeasible, EBAT uses a hybrid model with the main part of the data stored off-chain. At this tier, the system automatically forms a single digital "evidence package". This package is a logically connected set of files that must be self-sufficient for the complete reproduction and analysis of the situation by an independent auditor.

Its structure can be represented in a standardised format, such as JSON or XML, including several key sections:

1. InputData. A section containing the complete input data used to make the decision. To maintain integrity and privacy, sensitive data can be hashed or pseudonymised, but their structure and values must be recorded.

2. ModelExecution. This section contains complete information about the AI model. This is not just a name, but the exact version (e.g., a Git commit), its unique identifier (e.g., a SHA-256 hash of the model's binary file), and a list of versions of key libraries (TensorFlow, PyTorch, scikit-learn) used during execution. This is absolutely critical for ensuring reproducibility, as even a minor change in a library version can affect the result.

3. DecisionOutcome. Contains both the concise conclusion of the system ("denied/approved") and possibly more detailed information, such as a confidence score (prediction probability).

4. Explanation. Here, the counterfactual explanation generated in Tier 1 is stored in full text format, as well as metadata about the generation process itself (which XAI algorithm was used, how long it took).

The formed evidence package, which can range from a few kilobytes to many megabytes, is uploaded to the IPFS. The choice of IPFS is fundamental. Unlike traditional centralised storage (e.g., Amazon S3), where the owner company can delete or change data, and access is via a variable link (URL), IPFS uses content addressing (Benet, 2014). The system computes a cryptographic hash of the entire package and uses this hash (CID – Content Identifier) as its unique and immutable address. This creates a powerful guarantee of integrity: any attempt to change even one byte in the evidence package (e.g., to forge an explanation) will result in a complete change of its CID. Thus, the link to this package that will be recorded on the blockchain will become invalid, instantly exposing the attempt to interfere. To guarantee constant data availability, the package must be "pinned" on several IPFS nodes, which prevents its accidental deletion.

Tier 3: Creation of the immutable "trust anchor" on a permissioned blockchain

This is the final tier, which ensures the cryptographic immutability, finality, and ordering of the entire process. After receiving the CID of the evidence package from IPFS, the system initiates a transaction on a permissioned blockchain. For this architecture, the use of the Hyperledger Fabric framework is proposed. As noted by developers and researchers, Fabric is ideally suited for corporate tasks due to its modular architecture, support for confidential channels (which allows different participants to see only relevant transactions), absence of a speculative cryptocurrency, and high performance compared to Proof-of-Work systems (Androulaki *et al*., 2018).

The transaction process in Fabric is multi-stage. A client application forms a transaction proposal and sends it for endorsement to endorsing nodes, which simulate the execution of the corresponding smart contract (chaincode). If the simulation is successful and the nodes reach an agreement, they sign the result and return it to the client. The client collects the endorsements and sends the final transaction to the ordering service, which guarantees a single order of transactions for the entire network. After this, transactions are grouped into blocks and sent to all nodes for validation and recording in their copies of the ledger.

The transaction recorded on the blockchain in the EBAT architecture is extremely lightweight and contains only cryptographic fingerprints, serving as a "trust

anchor". Its structure can be implemented as a call to a chaincode function, for example CreateAuditTrail, with the following arguments:

- timestamp: the exact timestamp of the event, provided by the ordering service;
- decisionID: a unique identifier for the decision, allowing it to be linked to other systems;
- modelHash: the hash of the model file (SHA-256), ensuring that the correct version of the model was used;
- inputHash: the hash of the input data to confirm their immutability;
- explanationCID: the key field the CID of the evidence package from IPFS, which is a cryptographic link to the full context.

Thus, an unbreakable, ordered chain of evidence is permanently recorded on the blockchain. Any attempt to change one of the elements (e.g., to claim that a different AI model was used) will be immediately detected, as the hash of the changed model will not match the one permanently recorded on the blockchain. This creates the highest level of trust and accountability necessary for legal practice.

As noted in the methodology, the EBAT architecture is not a purely technical solution; its design was purposefully shaped by fundamental legal standards governing the handling of digital evidence. An analysis of two key documents the UK's "ACPO good practice guide for digital evidence" (2012) and the U.S. Federal Rules of Evidence (The U.S. Congress, 2024) allowed for the formulation of a set of engineering criteria that any system aiming to create legally significant logs must meet. The "ACPO good practice guide for digital evidence" establishes four core principles that serve as the gold standard in digital forensics. Principle 1 states that no action should change data that may be relied upon in court. Principle 3 requires that a full audit trail of all processes applied to digital evidence be created and preserved, in such a way that a third party can repeat those processes and achieve the same result. Together, these principles form the requirement for technical integrity and reproducibility. On the other hand, the U.S. Federal Rules of Evidence, particularly Rule 901, require the authentication of evidence. This means the proponent must produce sufficient evidence to support a finding that the item is what the proponent claims it is. In a digital context, this translates to proving that a log file, screenshot, or other artifact is not a forgery and was created by the specific process and at the specific time alleged. Based on these legal norms, the following key criteria for the design of the EBAT architecture were formulated four criterions:

Criterion 1: guaranteed data immutability. The system must cryptographically guarantee that, once recorded, no component of the evidence package (input data, model, explanation) can be altered without detection.

Criterion 2: process integrity and reproducibility. The system must record not only the result but the entire chain of actions, and do so in enough detail that an independent auditor can fully reproduce and verify the decision-making process.

Criterion 3: evidentiary authentication. The system must create an irrefutable proof of the log's origin and time of creation, linking a specific decision to specific data, a specific model, and a specific timestamp.

Criterion 4: semantic clarity. The record must be not only technically sound but also understandable to non-technical participants in the legal process (judges, lawyers), meaning it must contain an interpretation, not just raw data.

The EBAT architecture was designed so that each of its tiers directly addresses these criteria. Compliance with Criterion 1 (Immutability) is ensured at Tiers 2 and 3. Storing the evidence package in IPFS guarantees that any change to the data will result in a change to its Content Identifier. Since this CID is permanently recorded on the blockchain at Tier 3, any discrepancy between the CID on the blockchain and the actual CID of an altered file will be instantly detected. This creates a dual layer of protection against tampering. Compliance with Criterion 2 (Integrity and Reproducibility) is achieved through the comprehensive structure of the "evidence package" at Tier 2. The inclusion of not only the input data but also the exact version of the AI model and its dependencies (libraries) is key to reproducibility. An independent auditor, possessing this package, can not only view the result but can re-run the exact same model on the exact same data to verify that the outcome is identical. Compliance with Criterion 3 (Authentication) is the primary function of Tier 3. The transaction in Hyperledger Fabric serves as a form of digital notarisation. It contains an accurate timestamp from the ordering service, cryptographic hashes of all key components (model, input data), and the package's CID. This record, endorsed by network participants and included in the immutable chain of blocks, constitutes powerful proof of authenticity that satisfies the requirements of FRE Rule 901. Compliance with Criterion 4 (Semantic Clarity) is the unique advantage realised at Tier 1. Unlike other systems, EBAT makes the generation of a counterfactual explanation a mandatory part of the process. By including this human-readable explanation in the evidence package, the architecture transforms the "black box" into a transparent process, the results of which can be meaningfully analysed in court, not just technically verified. Thus, the EBAT architecture does not merely use blockchain as a technology for recording hashes; it is a holistic system designed "from law to code", where every architectural choice is justified by the specific requirements of legal evidentiary standards.

The proposed EBAT architecture is a direct response to the gaps identified in the literature review, offering a synthesised solution where existing research often focuses on separate components of the problem. Its effectiveness can be best understood by comparing it to the current scientific discourse in both the blockchain and Explainable AI domains. The foundation of the EBAT architecture rests on the principles of blockchain technology, which are extensively covered in the literature. The work of Y. Yuan & F. Wang (2019) provided a broad overview of blockchain

models and applications, establishing the technology as a robust framework for creating decentralised trust and ensuring data integrity through cryptographic linkage. The architecture builds upon this general model, but in a highly specialised manner. EBAT applied these principles to the niche but critical task of creating legally admissible evidence. This aligns with the vision outlined by K. Salah *et al*. (2019), who identified the potential for blockchain to bring accountability to AI but also highlighted the challenges of scalability and privacy. The EBAT architecture directly addresses these challenges through its hybrid on-chain/off-chain design, using Hyperledger Fabric for efficient, permissioned transactions and IPFS for scalable off-chain storage. This design choice is further validated by recent specialised applications. For instance, the auditing scheme for educational data proposed by F. Yu *et al*. (2024) demonstrated a modern blockchain application designed for trusted data detection. However, their model, while effective for verifying data integrity, exemplifies the very gap EBAT aims to fill: it ensures that the educational record is untampered but does not provide any mechanism to explain why an AI might have made a certain assessment based on that data. EBAT, in contrast, considers the integrity of the record and the intelligibility of its content to be equally important. This brings to the second pillar of our architecture: Explainable AI.

The fundamental challenge that XAI seeks to address is thoroughly documented in the comprehensive survey by A. Adadi & M. Berrada (2018). They provided a detailed taxonomy of the "black box" problem across various AI models and survey the landscape of explanation techniques designed to "peek inside". Their work clarified that there is no one-size-fits-all solution for explainability; the choice of method is highly context-dependent. The EBAT architecture acknowledges this by deliberately selecting a specific type of explanation-counterfactuals-based on their unique suitability for the legal domain. This choice is strongly supported by the legal and ethical analysis of S. Wachter *et al*. (2018), who argued that counterfactuals ("the outcome would have been different if…") provided actionable recourse and directly align with the principles of the GDPR's "right to explanation". While other techniques might provide technical insights, counterfactuals offer a narrative that is intuitively understandable to judges, lawyers, and the individuals affected by a decision. This aligns with the human-centric perspective advocated by G. Vilone & L. Longo (2021), who argued that the quality of an explanation should be judged not by its technical elegance but by its usefulness and comprehensibility to the end-user.

Thus, the primary contribution of the EBAT architecture to the scientific discourse is its synergistic synthesis. It moves beyond the siloed approaches prevalent in the literature. Unlike blockchain frameworks that focus solely on data integrity (like the one proposed by F. Yu *et al*. (2024)), and unlike theoretical XAI models that lack a secure, immutable storage mechanism for the explanations they generate (as is common in the works surveyed by A. Adadi & M. Berrada (2018)), EBAT binds the explanation to the record in a cryptographically inseparable manner. It transforms the blockchain from a simple timestamping service for opaque data into a permanent, verifiable ledger of justified decisions. By doing so, it provides a tangible engineering solution that addresses the complex socio-technical problem of building trust in AI within high-stakes, adversarial environments like the justice system.

Conclusions

This study conducted a comprehensive analysis of the accountability problem of artificial intelligence systems within the legal context, identifying a key gap between existing technical solutions that ensure the immutability of logs via blockchain and the fundamental legal requirement for their interpretation. In response to this challenge, a new hybrid architecture, the Explainable Blockchain Audit Trail, was developed and theoretically substantiated. The work provided a detailed description of its three-tiered structure, which synergistically combines Explainable AI methods, decentralised storage, and a permissioned blockchain with the goal of creating technically robust and legally significant evidence that complies with modern legal standards.

The conducted research allowed to draw several important conclusions. Firstly, the rapid implementation of artificial intelligence systems in the justice system and other regulated fields creates an acute need for mechanisms that ensure their transparency, accountability, and trust. Existing logging approaches, even with the use of blockchain technology, are insufficient as they only guarantee the technical immutability of records but do not solve the fundamental "black box" problem, leaving the logic of decisions opaque to lawyers and citizens. Secondly, for an audit trail to have real evidentiary value in court, it must meet not only technical criteria for integrity but also legal requirements for interpretation and comprehensibility. This means that the system must record not only the fact of a decision but also its justification in a human-accessible form.

The proposed architecture EBAT is a comprehensive solution that eliminates the identified gap between the technical and legal components of an audit. Through the synergistic integration of three key components a module for generating counterfactual explanations, the decentralised storage IPFS, and the permissioned blockchain Hyperledger Fabric-the EBAT architecture creates an audit trail that is simultaneously immutable, complete, reproducible, and, most importantly, legally significant. This allows for a transition from passive data recording to an active accountability system where every AI decision can be effectively verified and challenged. Thus, the research has achieved its goal by proposing an innovative engineering concept that has significant potential for building trust in automated systems in critically important areas.

Prospects for further research are primarily focused on the practical implementation and experimental validation of the proposed EBAT architecture. Creating a software prototype and testing it in realistic scenarios, such as credit

scoring or legal document analysis, is a necessary next step to assess its real-world performance, security, and cost-effectiveness. The results of such empirical validation would, in turn, provide the crucial foundation for the second priority area: the development of industry standards for the format and content of legally significant explanations of AI decisions, which would facilitate their unification and interoperability across different systems and jurisdictions.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160. doi: 10.1109/ACCESS.2018.2870052.

[2] Androulaki, E., *et al*. (2018). Hyperledger fabric: A distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth EuroSys conference* (pp. 1-15). New York: ACM. doi: 10.1145/3190508.3190538.

[3] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, article number 101805. doi: 10.1016/j.inffus.2023.101805.

[4] ACPO good practice guide for digital evidence (Version 5). (2023). Retrieved from https://www.digital-detective.net/digital-forensics-documents/ACPO_Good_Practice_Guide_for_Digital_Evidence_v5.pdf.

[5] Benet, J. (2014). IPFS - content addressed, versioned, P2P file system. *ArXiv*. doi: 10.48550/arXiv.1407.3561.

[6] Bharati, R., Khodke, P., Khadilkar, C., & Bawiskar, S. (2024). Forensic bytes: Admissibility and challenges of digital evidence in legal proceedings. *International Journal of Scientific Research in Science and Technology*, 11(16), 24-35. doi: 10.2139/ssrn.4896874.

[7] Casey, E. (Ed.) (2011). *Digital evidence and computer crime: Forensic science, computers, and the internet* (3rd ed.). Amsterdam: Academic Press.

[8] Faruk, M., Shahriar, H., Saha, B., & Barek, A. (2023). Security in electronic health records system: Blockchain-based framework to protect data integrity. In Y. Maleh, M. Alazab & I. Romdhani (Eds.), *Blockchain for cybersecurity in cyber-physical systems. Advances in information security* (Vol. 102, pp. 125-137). Cham: Springer. doi: 10.1007/978-3-031-25506-9_7.

[9] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5), 1-42. doi: 10.1145/3236009.

[10] Liu, Y., Wang, J., Yan, Z., Wan, Z., & Jäntti, R. (2023). A survey on blockchain-based trust management for Internet of Things. *IEEE Internet of Things Journal*, 10(7), 5898-5922. doi: 10.1109/JIOT.2023.3237893.

[11] Ramos, S., & Ellul, J. (2024). Blockchain for Artificial Intelligence (AI): Enhancing compliance with the EU AI Act through distributed ledger technology. A cybersecurity perspective. *International Cybersecurity Law Review*, 5, 1-20. doi: 10.1365/s43439-023-00107-9.

[12] Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). New York: ACM. doi: 10.1145/2939672.2939778.

[13] Salah, K., Rehman, M., Nizamuddin, N., & Al-Fuqaha, A. (2019). Blockchain for AI: Review and open research challenges. *IEEE Access*, 7, 10127-10149. doi: 10.1109/ACCESS.2018.2890507.

[14] Sutton, A., & Samavi, R. (2018). Tamper-proof privacy auditing for artificial intelligence systems. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18)* (pp. 5374-5378). Stockholm: IJCAI. doi: 10.24963/ijcai.2018/756.

[15] The U.S. Congress. (2024). *Federal rules of evidence*. Retrieved from https://www.uscourts.gov/file/78325/download.

[16] Verma, S. (2019). Weapons of math destruction: How big data increases inequality and threatens democracy. *The Journal for Decision Makers*, 44(2), 97-98. doi: 10.1177/0256090919853933.

[17] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89-106. doi: 10.1016/j.inffus.2021.05.009.

[18] Wang, X., Wu, Y.C., & Ma, Z. (2024). Blockchain in the courtroom: Exploring its evidentiary significance and procedural implications in U.S. judicial processes. *Frontiers in Blockchain*, 7, article number 1306058. doi: 10.3389/fbloc.2024.1306058.

[19] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *ArXiv*. doi: 10.48550/arXiv.1711.00399.

[20] Yuan, Y., & Wang, F. (2019). Blockchain and cryptocurrencies: Model, techniques, and applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(9), 1421-1428. doi: 10.1109/TSMC.2018.2854904.

[21] Yu, F., Lu, Q., Meng, L., Peng, J., Xi, J., & Li, X. (2024). A blockchain-based auditing scheme for educational data supporting trusted detection. In *Twelfth international conference on advanced cloud and big data (CBD)* (pp. 184-189). Brisbane: IEEE. doi: 10.1109/CBD65573.2024.00042.

[22] Zhang, Y., Tang, Y., Zhang, Z., Li, M., Li, Z., Khan, S., Chen, H., & Cheng, G. (2023). Blockchain-based practical and privacy-preserving federated learning with verifiable fairness. *Mathematics*, 11(5), article number 1091. doi: 10.3390/math11051091.

Архітектура EBAT: пояснюваний блокчейн для юридичного аудиту ШІ

Олексій Шамов

Дослідник інтелектуальних систем
Громадська організація «Освітня гільдія прав людини»
18010, вул. Різдвяна, 40/28, м. Черкаси, Україна
http://orcid.org/0009-0009-5001-0526

Анотація. Інтеграція штучного інтелекту у сфери з високим рівнем відповідальності, як-от правосуддя, створює «проблему чорної скриньки», де непрозорість алгоритмів підриває фундаментальні правові принципи, а наявні рішення для аудиту на основі блокчейну не здатні подолати критичний розрив між технічною цілісністю запису та його цінністю як юридичного доказу, що піддається інтерпретації. Це дослідження мало на меті розробити та теоретично обґрунтувати нову архітектуру системи аудиту, яка синергетично поєднує криптографічну надійність блокчейну з інтерпретаційною потужністю пояснюваного штучного інтелекту (ШІ) для створення логів рішень, що є не лише незмінними, але й юридично значущими та зрозумілими для людини. Методологія включала системний аналіз та синтез, огляд публікацій з наукометричних баз даних, аналіз правових стандартів для цифрових доказів та методи концептуального архітектурного проектування інформаційних систем. У дослідженні запропонована нова гібридна архітектура «Explainable Blockchain Audit Trail», спеціально розроблену для розв'язання цієї проблеми. Її новизна полягає у трирівневій структурі, яка, по-перше, вимагає обов'язкової генерації зрозумілих для людини контрфактичних пояснень для кожного рішення ШІ. По-друге, повний та самодостатній пакет доказів, що містить вхідні дані, специфікації моделі та згенероване пояснення, надійно зберігається у децентралізованому off-chain сховищі для гарантування його цілісності та доступності. Третій рівень створює незмінний «якір довіри» для цього пакету у приватному блокчейні, криптографічно пов'язуючи всі компоненти та забезпечуючи постійний, захищений від втручання запис про подію. Ця комплексна модель забезпечує повну відтворюваність процесу прийняття рішень та створює надійну, об'єктивну основу для судового перегляду та апеляції. Запропонована архітектура надає ключову теоретичну основу для розробки практичних інструментів для суддів, адвокатів та регуляторів, кінцевою метою якої є підвищення прозорості та захист прав громадян в епоху алгоритмічного прийняття рішень шляхом надання конкретних механізмів для оскарження непрозорих висновків

Ключові слова: юридичні технології; аудиторський слід; цифрові докази; незмінність логів; правосуддя