

EXPLAINABLE AI FOR TRUSTWORTHY DECISION SUPPORT IN DISTRIBUTED CLOUD AND EDGE COMPUTING: A CONCEPTUAL FRAMEWORK AND RESEARCH AGENDA

DR. ANIL KARADWAL
ASSISTANT PROFESSOR
ARIHANT COLLEGE, INDORE

Abstract

Artificial Intelligence (AI) is increasingly used to support automated decision-making in distributed cloud and edge computing environments. These environments involve multiple interconnected devices, servers, and data-processing nodes that collaborate to deliver intelligent services. However, the complexity of distributed infrastructure, the use of opaque machine-learning models, and the absence of consistent explanations can make AI-generated decisions difficult to understand, verify, and trust. This paper examines the role of Explainable Artificial Intelligence (XAI) in improving the transparency, accountability, and reliability of decision-support systems operating across distributed computing environments. Based on a conceptual review of relevant literature and trustworthy-AI guidance, the paper proposes a framework integrating explainability, distributed monitoring, privacy protection, security, human oversight, and accountability. The framework identifies important implementation challenges, including explanation consistency across nodes, communication overhead, model drift, data heterogeneity, and the trade-off between interpretability and predictive performance. The paper further outlines a research agenda for evaluating explainable distributed AI systems through measurable technical and human-centred criteria. It argues that explainability should be integrated throughout the AI lifecycle rather than added as a separate feature after deployment. The proposed framework provides a foundation for future empirical investigations into trustworthy AI-based decision support in cloud and edge computing.

Keywords: Explainable Artificial Intelligence, Distributed Systems, Cloud Computing, Edge Computing, Trustworthy AI, Decision Support, AI Governance, Model Interpretability.

1. Introduction

Artificial Intelligence has become an important component of modern computing infrastructure. Organisations increasingly employ machine-learning models to detect anomalies, allocate computational resources, identify cybersecurity threats, predict system failures, and optimise service delivery. These applications frequently operate in distributed environments where data processing and decision-making take place across multiple interconnected computing nodes.



Cloud computing provides scalable computational resources, while edge computing enables data processing closer to the point of generation. Their integration supports applications that require timely responses, including industrial automation, smart cities, connected healthcare infrastructure, and intelligent transportation systems. However, distributing AI models across different nodes introduces challenges involving data consistency, communication delays, security, and coordinated decision-making.

A further concern is the limited transparency of many advanced AI models. A system may generate a recommendation or initiate an automated action without providing sufficient information about the factors influencing its decision. Consequently, system administrators, developers, and other stakeholders may find it difficult to determine whether a decision is appropriate, reliable, or affected by inaccurate data.

Explainable Artificial Intelligence seeks to make AI behaviour and outputs understandable to relevant users. Research on XAI has explored techniques for explaining model predictions, comparing interpretable and complex models, and improving the transparency of AI-based decision-support systems (Schwalbe & Finzel, 2023; Explainable AI-Based Decision Support Systems, 2024).

The National Institute of Standards and Technology identifies explainability and interpretability as important characteristics of trustworthy AI, alongside reliability, safety, security, accountability, privacy, and fairness (Tabassi, 2023). These characteristics are particularly relevant to distributed environments because decisions may depend on data and models operating across different locations and administrative boundaries.

This paper therefore examines how explainability can be integrated into distributed cloud and edge computing to support more transparent, accountable, and reliable decisions.

2. Statement of the Problem

Traditional AI-based decision-support systems may provide predictions without communicating the reasoning or evidence behind them. In distributed computing environments, this limitation becomes more complex because the information contributing to a decision may originate from multiple nodes, datasets, and models.

For example, an edge device may identify unusual network activity, while a cloud-based model evaluates the same event using broader historical information. If the resulting assessments differ, administrators need to understand the factors responsible for the disagreement.

Several interconnected problems arise:

- **Limited interpretability:** Complex models may generate predictions that are difficult for users to understand.
- **Distributed accountability:** It may be unclear which model, node, or data source contributed to a particular decision.
- **Privacy and security risks:** Sharing detailed explanations across nodes may expose sensitive information.
- **Inconsistent explanations:** Different models or nodes may produce explanations that are difficult to compare.
-
- **Operational constraints:** Explanation generation can consume additional processing power, memory, and network bandwidth.

The central problem addressed in this paper is how explainability can be incorporated into distributed AI decision-support systems without disregarding computational efficiency, privacy, security, and operational reliability.

3. Objectives of the Study

The objectives of the paper are:

- To examine the relevance of Explainable AI in distributed cloud and edge computing.
- To identify the principal technical and governance challenges affecting trustworthy AI-based decision-making.
- To propose a conceptual framework for integrating explainability with monitoring, security, privacy, and human oversight.
- To identify measurable criteria for evaluating explainable distributed AI systems.
- To suggest future research directions for developing reliable and accountable AI-enabled distributed infrastructure.

4. Research Questions

The paper addresses the following research questions:

- **RQ1:** How can explainability improve the transparency and accountability of decisions generated in distributed computing environments?
- **RQ2:** What technical challenges arise when explanations must be generated and communicated across multiple computing nodes?
- **RQ3:** How can explainability be balanced with privacy, security, computational efficiency, and predictive performance?
- **RQ4:** Which evaluation measures are appropriate for assessing the effectiveness of explainable AI-based decision-support systems?

5. Research Methodology

This paper adopts a conceptual review methodology based on relevant published literature and established guidance on explainable and trustworthy AI. It uses the NIST Artificial Intelligence Risk Management Framework and selected scholarly reviews of explainability as its principal conceptual foundations.

The discussion is organised around five dimensions: explanation quality, distributed coordination, privacy and security, operational performance, and human oversight. These dimensions are synthesised into a proposed framework intended to guide system design and future empirical research.

The paper does not claim to be a systematic literature review, nor does it report primary data, experimental results, or statistical testing. The proposed framework and evaluation criteria should therefore be understood as conceptual contributions requiring subsequent validation.

6. Review of Literature

6.1 Explainable Artificial Intelligence

Explainable AI encompasses methods intended to make model behaviour and predictions more understandable. Some methods use inherently interpretable models, whereas others explain individual predictions produced by complex models.

Local explanation methods focus on a particular prediction, while global explanation methods attempt to describe broader model behaviour. Feature attribution, interpretable rule-based models, counterfactual explanations, and visualisation techniques are examples of approaches used for different explanatory purposes.

Schwalbe and Finzel (2023) provide a comprehensive taxonomy of XAI methods and concepts, illustrating the diversity of available approaches. Their work supports the need to select explanation techniques according to the model, task, and intended user.

However, an explanation should not automatically be interpreted as proof that a decision is correct. Explanations may be incomplete, unstable, or misleading if the method does not faithfully represent the model's behaviour.

6.2 XAI in Decision-Support Systems

AI-based decision-support systems assist users by analysing information and generating predictions, recommendations, or classifications. Explainability can help users understand the evidence behind a recommendation and determine whether further investigation is necessary.

A 2024 review of explainable AI-based decision-support systems discusses methods, applications, challenges, and future research opportunities in this area. Such research

highlights the importance of providing explanations that are useful to the people responsible for interpreting and acting on AI outputs.

In distributed systems, decision support requires additional consideration because information may be generated and processed by different components. An explanation that describes only the final prediction may be insufficient when the decision depends on several nodes or models.

6.3 Trustworthy AI and Governance

The NIST AI Risk Management Framework identifies validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness as characteristics relevant to trustworthy AI (Tabassi, 2023).

The framework also emphasises that trustworthy AI requires consideration of the system's context and the relationships among different risk-management objectives. Improving explainability alone does not guarantee that an AI system is trustworthy.

For distributed AI, these principles imply that explanations should be considered alongside data quality, access control, system monitoring, privacy safeguards, and mechanisms for reviewing decisions.

6.4 Research Gap

The reviewed literature establishes the importance of explainability and trustworthy AI, but the integration of these principles into distributed cloud and edge decision-making requires further investigation.

A key research opportunity is to examine how explanations can remain understandable and sufficiently consistent when decisions depend on multiple models, heterogeneous datasets, and nodes with different computational capabilities.

This paper addresses that opportunity by proposing a conceptual framework linking explanation generation to distributed coordination, risk management, monitoring, and human oversight. This is a proposed research direction rather than a claim that existing literature has failed to address every component individually.

7. Proposed Conceptual Framework

The proposed framework, termed the **Explainable Distributed Decision Support Framework (EDDSF)**, organises trustworthy AI-based decision-making into six interacting components.

7.1 Data and Context Layer



This layer collects information from edge devices, cloud platforms, sensors, applications, and other authorised sources. Data provenance, quality checks, timestamps, and source identification help establish the context in which a decision is generated.

The objective is to ensure that explanations can refer to relevant evidence without exposing unnecessary sensitive information.

7.2 AI Decision Layer

Machine-learning models analyse the available information and produce predictions, classifications, or recommendations. Depending on the application, processing may occur locally at an edge node, centrally in the cloud, or across multiple nodes.

Model selection should consider predictive performance, computational requirements, interpretability, and the consequences of incorrect decisions.

7.3 Explanation Layer

This layer generates explanations appropriate to the model and intended audience. Possible techniques include feature attribution, interpretable rules, counterfactual explanations, and model-level summaries.

For example, a system detecting a possible network intrusion might explain that an alert was influenced by an unusual connection frequency, an unexpected traffic pattern, and a deviation from established behaviour.

The explanation should identify the relevant evidence and communicate uncertainty where appropriate. It should not imply certainty when the underlying model provides only a probabilistic prediction.

7.4 Distributed Coordination Layer

This component coordinates explanations generated by different nodes. It records which model and data sources contributed to a decision and helps reconcile differences between local and central assessments.

A distributed explanation record could contain:

- Decision identifier and timestamp.
- Model and node identifiers.
- Relevant input-data references.
- Generated prediction and confidence information.
- Explanation method and principal factors.
- Model or data version.

- Any human review or subsequent correction.
- The record should be designed to minimise unnecessary data sharing and protect sensitive operational information.

7.5 Trust, Security, and Governance Layer

This layer addresses authentication, access control, privacy protection, auditability, fairness checks, and security monitoring. It also defines the circumstances under which human review is required.

High-impact decisions should have stronger review and accountability mechanisms than routine, low-risk operations. The specific safeguards should depend on the application and the potential consequences of failure.

7.6 Human Oversight and Feedback Layer

System administrators, analysts, and other authorised users review explanations, investigate anomalies, and provide feedback when decisions appear inappropriate.

Feedback can inform model improvement, policy updates, and monitoring procedures. However, feedback should be checked for quality and potential bias before being incorporated into future training or operational decisions.

The six components form a continuous process: data are collected, decisions are generated, explanations are produced, distributed contributions are coordinated, risks are monitored, and human feedback informs subsequent evaluation.

8. Proposed Evaluation Framework

The framework can be evaluated in future experiments using representative distributed AI workloads. The following measures are proposed for this purpose.

Evaluation dimension	Suggested measures	Purpose
Explanation quality	Fidelity, stability, comprehensibility	Assess whether explanations represent model behaviour and are understandable
Predictive performance	Accuracy, precision, recall, F1-score	Evaluate decision quality for the selected task
Operational efficiency	Explanation latency, CPU and memory use	Measure computational overhead
Network efficiency	Additional communication volume, transmission latency	Assess the cost of exchanging explanation information

Privacy and security	Information leakage tests, access-control violations	Examine exposure and unauthorised access risks
Accountability	Audit completeness, decision traceability	Determine whether decisions can be reconstructed and reviewed
Human oversight	Review time, error detection, appropriate reliance	Assess the practical usefulness of explanations

These measures are proposed evaluation criteria, not experimental findings. Their suitability will depend on the application, the model architecture, the operational environment, and the risks associated with incorrect decisions.

Future experiments could compare three configurations: a distributed AI system without explanation facilities, a system with locally generated explanations, and a system combining local explanations with distributed coordination and governance.

Experiments should use the same underlying workload and clearly specified test conditions. They should examine whether the proposed coordination mechanisms improve traceability and explanation usefulness, and whether any improvement introduces unacceptable computational or privacy costs.

9. Discussion

The proposed framework suggests that explainability should be treated as an operational capability rather than merely a visual presentation of model outputs.

First, distributed decision-making requires traceability. When multiple models contribute to an outcome, users need to understand which components influenced the final decision. Explanation records and model-version information may support this requirement.

Second, explanations must be adapted to their audience. A network administrator may need technical details about features, model versions, and system behaviour, whereas a manager may require a concise account of the principal risks and the recommended response. An explanation that is technically detailed but difficult to interpret may not support effective decision-making.

Third, explainability creates potential trade-offs. Detailed explanation records can increase storage and communication requirements. Sharing feature-level information may reveal sensitive characteristics of users, systems, or infrastructure. Furthermore, a plausible explanation may still be inaccurate if it does not faithfully represent the model's decision process.



Fourth, trust should not be equated with unquestioning acceptance. An effective explanation should enable appropriate reliance, including the ability to question, override, or escalate a recommendation when the evidence is insufficient.

Finally, governance must continue after deployment. Changes in data distributions, operating conditions, model versions, and cyber threats can affect both prediction quality and explanation reliability. Continuous monitoring and periodic review are therefore important elements of a trustworthy distributed AI system.

10. Practical Applications

The framework may be relevant to several application areas.

- **Cybersecurity:** Explainable anomaly detection can help security analysts understand why a network event has been flagged and prioritise investigation.
- **Smart manufacturing:** Distributed models can explain potential equipment failures using relevant sensor patterns, supporting maintenance planning.
- **Healthcare infrastructure:** Explainable systems can assist authorised professionals in interpreting alerts generated across connected devices and clinical computing systems. Such systems require strong privacy safeguards and appropriate professional oversight.
- **Smart cities:** Distributed traffic-management systems can communicate the factors influencing congestion forecasts or traffic-control recommendations.
- **Cloud resource management:** Explainable scheduling systems can identify factors affecting workload placement, resource allocation, or service-level risk.

These applications illustrate possible uses of the proposed framework. Their actual effectiveness would need to be established through domain-specific evaluation.

11. Limitations and Future Research Directions

This paper has several limitations. It is conceptual and does not provide primary empirical evidence, benchmark comparisons, or quantitative validation of the proposed framework. Its recommendations must therefore be tested before claims about performance improvements can be made.

Future research could investigate the following questions:

- How can explanations from heterogeneous models be represented in a common, comparable format?
- Which explanation methods provide the best balance between fidelity, usability, and computational cost in edge environments?
- How can privacy-preserving explanation sharing be implemented without substantially reducing explanatory value?

- How can systems detect explanations that become unreliable after model updates or data drift?
- How do technical expertise and task risk influence human reliance on explanations?
- Can standardised explanation records improve accountability across organisations that operate different parts of a distributed system?

A useful next step would be to implement a prototype and evaluate it on a documented distributed workload. Such research should define the dataset, baseline systems, explanation techniques, performance measures, and experimental procedures before drawing conclusions.

12. Conclusion

Explainable Artificial Intelligence offers a promising approach to improving the transparency and accountability of AI-based decision support in distributed cloud and edge computing. However, the distributed nature of these environments introduces additional requirements concerning explanation coordination, data provenance, privacy, security, and operational efficiency.

This paper proposed the Explainable Distributed Decision Support Framework, integrating data context, AI decision-making, explanation generation, distributed coordination, governance, and human oversight. It also outlined evaluation measures and future research questions for assessing the framework.

The principal contribution is a structured conceptual approach for examining explainability as part of the wider architecture and governance of distributed AI systems. The framework is not yet empirically validated, and its effectiveness remains a question for future research. Ultimately, trustworthy distributed AI requires more than understandable predictions: it requires explanations that are faithful, contextually appropriate, secure, auditable, and useful for informed human decision-making.

References

- Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. (2024). *Electronics*, 13(14), Article 2842. <https://doi.org/10.3390/electronics13142842>
- Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*. <https://doi.org/10.1007/s10618-022-00867-8>
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>

- DARPA. (n.d.). *Explainable Artificial Intelligence (XAI)*. Defense Advanced Research Projects Agency.
<https://www.darpa.mil/research/programs/explainable-artificial-intelligence>