The American Journal of Engineering and Technology

ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO. 206-211

Toward Trustworthy and Domain-Transcendent Explainable Artificial Intelligence: A Unified Theoretical and Applied Framework Across Healthcare, Finance, Energy, Engineering, and Organizational Systems

Dr. Elias Morgenstern

Department of Computational Intelligence and Decision Sciences

University of Applied Sciences Zurich Switzerland

OPEN ACCESS

SUBMITTED 01 October 2025
ACCEPTED 15 October 2025
PUBLISHED 31 October 2025
VOLUME Vol.07 Issue 10 2025

CITATION

Dr. Elias Morgenstern. (2025). Toward Trustworthy and Domain-Transcendent Explainable Artificial Intelligence: A Unified Theoretical and Applied Framework Across Healthcare, Finance, Energy, Engineering, and Organizational Systems. The American Journal of Engineering and Technology, 7(10), 206–211. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/7175

Abstract

Background: Artificial intelligence has achieved unprecedented predictive and decision-making capabilities across diverse domains such as healthcare, finance, energy systems, civil engineering, and organizational management. However, the increasing opacity of complex machine learning and deep learning models has raised critical concerns regarding trust, accountability, fairness, and regulatory compliance. Explainable Artificial Intelligence (XAI) has emerged as a pivotal paradigm aimed at addressing these concerns by rendering AI systems transparent, interpretable, and human-understandable.

Objective: This research develops a comprehensive, domain-transcendent theoretical and applied framework for explainable artificial intelligence by synthesizing insights from multidisciplinary applications including medical diagnostics, financial risk management, energy forecasting, structural engineering, organizational agility prediction, and counterfactual reasoning. The study seeks to identify unifying principles, methodological patterns, and conceptual gaps that limit the scalability and reliability of XAI systems across real-world settings.

Methods: A qualitative, theory-driven research methodology is employed, grounded strictly in an in-depth analytical synthesis of contemporary peer-reviewed literature on XAI. The methodology integrates interpretability taxonomies, post-hoc and intrinsic explanation strategies, counterfactual reasoning mechanisms, and self-explainable model architectures. Emphasis is placed on descriptive methodological reasoning rather than mathematical formalization, aligning with interdisciplinary accessibility requirements.

Results: The findings reveal that while XAI techniques demonstrate significant domain-specific effectiveness, they remain fragmented in conceptual alignment and evaluation standards. Medical and biological applications prioritize causal and feature-attribution explanations, finance emphasizes transparency and regulatory compliance, energy systems focus on temporal explainability, and engineering domains demand structural logic validation. A unifying theoretical scaffold based on explanation purpose, stakeholder cognition, and decision risk is identified.

Conclusion: The study concludes that future progress in XAI depends on transitioning from tool-centric explanations to cognition-aware, context-sensitive, and ethically grounded explanatory ecosystems. The proposed unified framework advances explainable AI beyond interpretability toward actionable trust, supporting responsible deployment across high-stakes domains.

Keywords: *Explainable Artificial Intelligence, Trustworthy AI, Interpretability, Counterfactual Explanations, Domain-Specific AI, Responsible AI*

Introduction

Artificial intelligence has transitioned from an experimental computational paradigm to a foundational infrastructure underpinning modern decision-making across scientific, industrial, and societal domains. From diagnosing complex diseases and forecasting renewable energy output to managing financial risk and optimizing organizational performance, AI-driven systems increasingly shape outcomes that directly affect human lives and institutional stability. Despite this transformative potential, the rapid adoption of advanced machine learning models has introduced a fundamental tension between predictive accuracy and interpretability. As models grow in complexity, particularly with the rise of deep neural networks and ensemble learning architectures, their internal reasoning processes become progressively opaque, giving rise to what is often described as the “black box” problem (Arrieta et al., 2019).

This opacity presents significant challenges. In healthcare, clinicians require transparent justification for diagnostic predictions to ensure patient safety and ethical accountability. In finance, regulatory bodies demand explainable credit decisions to prevent discrimination and systemic risk. In energy systems, operators need interpretable forecasts to manage infrastructure reliability, while in engineering and organizational contexts, decision-makers must understand AI-driven recommendations to ensure alignment with physical constraints and strategic objectives (Kuzlu et al., 2020; Nayak, 2022; Shafiabady et al., 2023). Consequently, explainability has evolved from a desirable feature into a critical requirement for the responsible deployment of artificial intelligence.

Explainable Artificial Intelligence (XAI) represents a multidisciplinary response to this challenge, encompassing methods, frameworks, and philosophies designed to make AI systems understandable to human stakeholders. Rather than focusing solely on performance metrics, XAI emphasizes transparency, interpretability, fairness, and trust. Early efforts in XAI were largely technical, concentrating on post-hoc explanation tools such as feature importance measures and visualization techniques. However, recent scholarship has expanded the scope of XAI to include human-centered design, ethical governance, and domain-specific interpretability requirements (Gunning et al., 2021; Gohel et al., 2021).

Despite substantial progress, the XAI landscape remains fragmented. Techniques effective in one domain often fail to generalize to others, and evaluation standards for explanations lack consistency. Moreover, many XAI methods prioritize developer-oriented explanations while neglecting the cognitive needs of end-users such as clinicians, regulators, or organizational leaders. This fragmentation underscores a critical literature gap: the absence of a unified, domain-transcendent framework that integrates theoretical principles with applied insights across diverse fields.

This article addresses this gap by developing a comprehensive, publication-ready synthesis of explainable artificial intelligence grounded strictly in existing scholarly literature. Drawing on applications in medical imaging, genomic analysis, financial risk prediction, energy forecasting, structural engineering, organizational agility assessment, and counterfactual reasoning, the study seeks to articulate common explanatory principles, methodological convergences, and unresolved challenges. By doing so, it advances a holistic understanding of XAI as not merely a set of tools, but as an evolving epistemological framework for trustworthy artificial intelligence.

Methodology

The methodological foundation of this research is qualitative, interpretive, and theory-driven, reflecting the conceptual and interdisciplinary nature of explainable artificial intelligence. Rather than employing empirical experimentation or quantitative modeling, the study adopts an integrative analytical synthesis approach. This methodology is particularly suitable for examining a rapidly evolving research field where conceptual clarity, theoretical coherence, and cross-domain applicability are paramount.

The research process began with a systematic examination of peer-reviewed journal articles, conference proceedings, and authoritative surveys focusing on explainable artificial intelligence and its applications across multiple domains. The selected literature spans healthcare, finance, energy systems, civil engineering, organizational science, and foundational XAI theory. Each source was analyzed in depth to extract its underlying assumptions about explainability, the methods employed, the stakeholders addressed, and the practical constraints encountered.

A central methodological principle guiding this study is domain contextualization. Instead of treating explainability as a monolithic concept, the analysis recognizes that explanations are inherently relational, shaped by the domain in which an AI system operates and the users who interact with it. For example, explainability in medical imaging prioritizes causal reasoning and clinical relevance, whereas explainability in financial risk management emphasizes transparency, auditability, and compliance (Houssein et al., 2025; Nayak, 2022). By systematically comparing these contextual requirements, the methodology uncovers both domain-specific nuances and cross-cutting explanatory patterns.

Another key methodological dimension involves categorizing XAI approaches into intrinsic and post-hoc methods, as established in the foundational literature (Arrieta et al., 2019). Intrinsic explainability refers to models that are interpretable by design, such as decision trees or rule-based systems. Post-hoc explainability involves techniques applied after model training to interpret complex models, including feature attribution methods, surrogate models, and counterfactual explanations. This distinction provides a conceptual scaffold for organizing the diverse methods discussed across the literature.

The methodology also integrates emerging perspectives on self-explainable and counterfactual AI systems. Self-explainable models embed explanation mechanisms directly into their architecture, enabling real-time interpretability without reliance on external tools (Hou et al., 2024). Counterfactual explanations, on the other hand, focus on minimal changes to input features that would alter a model's decision, offering intuitive "what-if" scenarios for users (You et al., 2023). These approaches are examined not only in terms of technical feasibility but also in relation to human cognition and decision-making processes.

Throughout the analysis, emphasis is placed on descriptive clarity rather than mathematical formalism. All algorithmic concepts, data transformations, and inferential mechanisms are explained through detailed narrative descriptions, ensuring accessibility to readers from diverse disciplinary backgrounds. This methodological choice aligns with the overarching objective of XAI itself: to make complex systems understandable without sacrificing rigor.

Results

The integrative analysis yields several significant findings that illuminate both the current state and the structural limitations of explainable artificial intelligence across domains. One of the most salient results is the observation that explainability is not a singular property of an AI system but a multidimensional construct shaped by purpose, audience, and risk context. This insight challenges simplistic interpretations of XAI as merely a technical add-on and underscores the need for a more nuanced conceptualization.

In healthcare and biomedical applications, explainability is closely tied to causality and biological plausibility.

Studies on medical imaging and gene biomarker identification demonstrate that clinicians value explanations that align with established physiological knowledge and support diagnostic reasoning (Yagin et al., 2023; Houssein et al., 2025). Feature attribution methods are widely used to highlight regions of medical images or genetic markers associated with disease outcomes. However, the results indicate that such explanations are only trusted when they correspond to clinically meaningful patterns rather than abstract statistical correlations.

In financial and supply chain contexts, the primary function of explainability is accountability. Financial risk prediction models must justify their decisions to regulators, auditors, and customers, particularly in high-stakes scenarios such as credit approval or fraud detection (Yi et al., 2023; Nayak, 2022). The analysis reveals that post-hoc explanation techniques, including feature importance rankings and counterfactual scenarios, are effective in enhancing transparency but often struggle to capture complex temporal and behavioral dynamics inherent in financial data.

Energy systems and engineering applications present a different explanatory emphasis. In solar power forecasting and structural engineering, explainability is valued for its ability to validate model predictions against physical laws and engineering intuition (Kuzlu et al., 2020; Saleh et al., 2023). The results show that XAI tools can enhance confidence in AI-driven forecasts and design recommendations, particularly when explanations reveal how environmental variables or structural parameters influence outcomes. However, the reliance on post-hoc explanations introduces challenges related to stability and consistency across different operational conditions.

Organizational and behavioral AI applications further expand the explanatory landscape. In predicting organizational agility or thermal comfort in buildings, explainability supports strategic planning and policy formulation by clarifying the relationships between human behavior, environmental factors, and performance outcomes (Ngarambe et al., 2020; Shafiabady et al., 2023). The findings suggest that in such contexts, explanations must balance analytical depth with interpretive simplicity to remain actionable for decision-makers.

Across all domains, a recurring result is the lack of standardized evaluation criteria for explanations. While predictive accuracy is easily quantified, the quality of explanations remains subjective and context-dependent. This absence of universal benchmarks limits the comparability of XAI methods and complicates their integration into regulatory and operational frameworks.

Discussion

The findings of this study invite a deeper reflection on the theoretical and practical implications of explainable artificial intelligence. One of the most profound insights emerging from the analysis is that explainability should be understood as a socio-technical phenomenon rather than a purely computational attribute. Explanations are meaningful only insofar as they resonate with human cognitive models, institutional norms, and ethical expectations. This perspective aligns with contemporary critiques of reductionist XAI approaches that equate interpretability with feature visualization alone (Gohel et al., 2021; Holzinger et al., 2020).

A critical theoretical implication concerns the relationship between explanation and trust. Trust in AI systems does not arise automatically from transparency; rather, it is mediated by users' prior knowledge, domain expertise, and perceived alignment between explanations and real-world experience. For example, a technically accurate explanation that contradicts a clinician's understanding of disease pathology may erode trust rather than enhance it. This underscores the importance of context-aware and user-centered explanation design, an area that remains underdeveloped in much of the current XAI literature.

The discussion also highlights inherent trade-offs between model complexity and interpretability. While intrinsic models offer clarity, they may lack the expressive power needed for complex tasks. Conversely, high-performing deep learning models often require post-hoc explanations that are approximate and potentially misleading. Counterfactual explanations offer a promising middle ground by focusing on decision boundaries rather than internal representations, yet they raise ethical concerns regarding feasibility and fairness when suggested changes are unrealistic or socially sensitive (You et al., 2023).

From a methodological standpoint, the fragmentation of XAI approaches across domains suggests the need for a unifying explanatory framework grounded in purpose rather than technique. Such a framework would begin by identifying the primary goal of explanation—whether it is accountability, validation, learning, or persuasion—

and then selecting methods aligned with that goal. This purpose-driven approach could mitigate the tendency to apply generic XAI tools without regard for contextual relevance.

The study also acknowledges several limitations. By relying exclusively on existing literature, the analysis does not incorporate empirical user studies that could provide direct evidence of explanation effectiveness. Additionally, the rapidly evolving nature of XAI means that new methods and paradigms may emerge beyond the scope of the reviewed sources. Nevertheless, the depth and breadth of the synthesis provide a robust foundation for future empirical and theoretical work.

Looking forward, future research should prioritize the development of standardized explanation evaluation metrics that account for human factors and domain-specific risks. Interdisciplinary collaboration between AI researchers, domain experts, ethicists, and policymakers will be essential to translate XAI principles into practical governance frameworks. Moreover, advances in self-explainable and cognitively inspired AI architectures hold promise for embedding interpretability directly into intelligent systems, reducing reliance on post-hoc approximations.

Conclusion

This research has presented an extensive, theory-driven exploration of explainable artificial intelligence as a foundational pillar of trustworthy and responsible AI deployment. By synthesizing insights from healthcare, finance, energy, engineering, organizational science, and foundational XAI theory, the study demonstrates that explainability is neither a universal solution nor a mere technical accessory. Instead, it is a context-sensitive, purpose-driven construct that must be carefully aligned with domain requirements, stakeholder expectations, and ethical considerations.

The central contribution of this article lies in articulating a unified, domain-transcendent perspective on XAI that moves beyond fragmented tool-based approaches. The analysis reveals that meaningful explanations emerge at the intersection of technical rigor, human cognition, and institutional accountability. As artificial intelligence continues to permeate high-stakes decision environments, the imperative for explainability will only intensify.

Ultimately, the future of AI depends not solely on its capacity to predict or optimize, but on its ability to justify, communicate, and align with human values. Explainable artificial intelligence, when thoughtfully designed and contextually grounded, offers a pathway toward this future—one in which intelligent systems are not only powerful, but also comprehensible, accountable, and worthy of trust.

References

1. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. BCAM – Basque Center for Applied Mathematics.

2. Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: Current status and future directions. arXiv.

3. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2(3).

4. Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. P., & Samek, W. (2020). xxAI – Beyond explainable AI. Springer.

5. Hou, J., Sicen, L., Bie, Y., Wang, H., Tan, A., Luo, L., & Chen, H. (2024). Self-explainable AI for medical image analysis: A survey and new outlooks. arXiv.

6. Houssein, E. H., Gamal, A. M., Younis, E. M. G., & Mohamed, E. (2025). Explainable artificial intelligence for medical imaging systems using deep learning: A comprehensive review. Cluster Computing, 28, 469.

7. Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö. (2020). Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools. IEEE Access, 8, 187814–187823.

8. Nayak, S. (2022). Harnessing explainable AI (XAI) for transparency in credit scoring and risk management in fintech. International Journal of Applied Engineering and Technology, 4, 214–236.

9. Ngarambe, J., Yun, G. Y., & Santamouris, M. (2020). The use of artificial intelligence methods in the prediction of thermal comfort in buildings: Energy implications of AI-based thermal comfort controls. Energy and Buildings, 211, 109807.

10. Saleh, M., AlHamaydeh, M., & Zakaria, M. (2023).

Shear capacity prediction for reinforced concrete deep beams with web openings using artificial intelligence methods. Engineering Structures, 280, 115675.

11. Shafiabady, N., Hadjinicolaou, N., Din, F. U., Bhandari, B., Wu, R. M. X., & Vakilian, J. (2023). Using artificial intelligence to predict organizational agility. PLoS ONE, 18, e0283066.

12. Yi, Z., Liang, Z., Xie, T., & Li, F. (2023). Financial risk prediction in supply chain finance based on buyer transaction behavior. Decision Support Systems, 170, 113964.

13. Yagin, F. H., Cicek, İ. B., Alkhateeb, A., Yagin, B., Colak, C., Azzeh, M., & Akbulut, S. (2023). Explainable artificial intelligence model for identifying COVID-19 gene biomarkers. Computers in Biology and Medicine, 154, 106619.

14. You, D., Niu, S., Dong, S., Yan, H., Chen, Z., Wu, D., Shen, L., & Wu, X. (2023). Counterfactual explanation generation with minimal feature boundary. Information Sciences, 625, 342–366.

15. Lundberg, S., Erion, G., & Lee, S. I. (2019). Explainable AI for trees: From local explanations to global understanding. arXiv.