

A REVIEW OF DYNAMIC TRUST-AWARE EXPLAINABILITY MODEL FOR ARTIFICIAL INTELLIGENCE SYSTEMS OPERATING IN HIGH-CONSEQUENCE DECISION DOMAINS

Sapna Singh¹, Mrs. Arifa Khan²

¹Master of Technology, Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

²Assistant Professor, Department of Computer Science and Engineering, Lucknow Institute of Technology, Lucknow, India

Abstract - Artificial Intelligence (AI) systems are increasingly deployed in high-consequence decision domains such as healthcare, autonomous transportation, finance, and defense, where erroneous or opaque decisions can result in severe societal, ethical, and economic impacts. In such environments, establishing calibrated human trust and ensuring model explainability are critical requirements. While substantial progress has been made in Explainable AI (XAI) and computational trust modeling independently, the integration of dynamic trust mechanisms with adaptive explainability remains fragmented across the literature. This review systematically analyzes existing approaches to trust-aware explainability, focusing on models that dynamically adjust explanations based on contextual risk, user expertise, and system performance. We categorize current research into intrinsic and post-hoc explainability methods, static and dynamic trust estimation frameworks, and integrated trust-explainability architectures. Furthermore, we evaluate domain-specific implementations and comparative evaluation metrics used to assess explanation quality and trust calibration. The review identifies key research gaps, including the absence of standardized benchmarks, limited real-time adaptability, and insufficient human-centered validation. Finally, we outline future research directions aimed at developing unified, context-sensitive, and regulatory-compliant trust-aware explainable AI systems for safety-critical applications.

Key Words: Explainable Artificial Intelligence (XAI), Dynamic Trust Modelin, Trust-Aware Systems, High-Consequence Decision Domains, Human-AI Interaction, AI Accountability

1. INTRODUCTION

Artificial Intelligence (AI) systems are increasingly embedded in socio-technical infrastructures where their decisions directly influence human safety, legal standing, and economic stability. In high-impact environments, performance accuracy alone is insufficient; systems must also provide intelligible reasoning and maintain calibrated trust relationships with users. The emergence of Explainable Artificial Intelligence (XAI) and computational trust modeling reflects the recognition that transparency, accountability, and reliability are foundational for

responsible AI deployment (Gunning, 2017; Doshi-Velez and Kim, 2017). This section contextualizes the review by examining the operational landscape of AI in critical domains, the theoretical need for dynamic trust-aware explainability, and the methodological approach adopted for literature synthesis.

1.1 Background

1.1.1 AI in Critical Sectors

AI technologies are now widely deployed in healthcare for diagnostic support and treatment planning, in autonomous vehicles for perception and navigation, in defense for surveillance and threat assessment, and in finance for credit scoring and fraud detection. In healthcare, deep learning models have achieved expert-level performance in medical imaging tasks, yet their opacity raises concerns about clinical accountability (Esteva et al., 2017). Similarly, autonomous driving systems rely on complex perception pipelines where failure can result in catastrophic consequences, highlighting the necessity for interpretable decision pathways (Bonneton, Shariff and Rahwan, 2016). In finance, algorithmic decision-making affects credit eligibility and risk profiling, often raising fairness and transparency issues (Barocas and Selbst, 2016). Across these sectors, AI systems operate under regulatory scrutiny and ethical constraints, reinforcing the demand for explainable and trustworthy models.

1.1.2 Importance of Trust and Explainability

Trust in AI systems is a multidimensional construct involving reliability, predictability, and perceived competence. Without appropriate explanations, users may either over-trust automated systems (automation bias) or under-trust them (algorithm aversion), both of which degrade decision quality (Lee and See, 2004; Dietvorst, Simmons and Massey, 2015). Explainability mechanisms aim to render model reasoning comprehensible, thereby supporting trust calibration rather than blind reliance. The DARPA XAI initiative formalized this need by emphasizing that AI systems must provide human-understandable justifications to support operational decision-making (Gunning, 2017). Consequently, trust and explainability are interdependent constructs in high-consequence domains.

1.2 Need for the Review

1.2.1 Need for Dynamic, Context-Aware Trust Metrics

Traditional trust models often assume static reliability estimates; however, AI systems deployed in dynamic environments experience distribution shifts, adversarial inputs, and contextual variability. Trust must therefore be adaptive, updating based on performance feedback, environmental uncertainty, and user interaction history. Bayesian and probabilistic trust models have been proposed to dynamically quantify system reliability under uncertainty (Yu, Singh and Sycara, 2004). In safety-critical domains, context-sensitive trust estimation is essential for real-time risk mitigation and human oversight. Despite progress in adaptive AI, integration between trust dynamics and explanation generation remains limited.

1.2.2 Limitations of Static Explainability Approaches

Many explainability techniques—such as post-hoc feature attribution methods—provide local explanations without accounting for user expertise, situational risk, or evolving model behavior (Ribeiro, Singh and Guestrin, 2016; Lundberg and Lee, 2017). These static explanations may be technically accurate yet cognitively misaligned with end-users. Moreover, explanation fidelity does not automatically translate into user trust, especially when explanations fail to adapt to contextual requirements. This limitation motivates the exploration of dynamic, trust-aware explainability frameworks that tailor explanatory depth and format according to operational context.

1.3 Scope and Objectives

1.3.1 Defining High-Consequence Decision Domains

High-consequence decision domains refer to environments in which AI-driven outcomes can result in significant physical harm, legal liability, ethical violations, or large-scale economic loss. Examples include clinical diagnostics, autonomous navigation, military operations, and financial risk assessment. These domains are characterized by high uncertainty, strict regulatory oversight, and the necessity for human accountability. The European Union's regulatory framework for trustworthy AI emphasizes transparency, robustness, and human oversight as essential requirements in such contexts (European Commission, 2019). Therefore, any explainability model designed for these domains must satisfy both technical robustness and socio-legal accountability.

1.3.2 Dynamic Trust-Aware Models Matter

Dynamic trust-aware explainability models aim to align system transparency with evolving risk levels and user needs. Unlike static frameworks, these models adjust explanatory granularity based on contextual signals such as anomaly detection, model confidence, or user feedback. This

adaptability supports calibrated trust, reduces cognitive overload, and enhances collaborative human-AI decision-making. By synthesizing research across XAI and trust modeling, this review seeks to identify conceptual gaps, methodological trends, and integration strategies necessary for next-generation AI systems operating in safety-critical environments.

1.4 Methodology of Literature Selection

1.4.1 Databases and Search Strategy

The literature for this review was systematically collected from major academic databases, including IEEE Xplore, Scopus, and Web of Science. These repositories were selected due to their comprehensive indexing of peer-reviewed journals and conference proceedings in artificial intelligence, human-computer interaction, and computational trust.

1.4.2 Inclusion and Exclusion Criteria

Included studies met the following criteria: (i) focus on explainable AI, trust modeling, or integrated frameworks; (ii) application in safety-critical or high-risk domains; and (iii) publication in peer-reviewed venues. Excluded works comprised non-peer-reviewed articles, purely opinion-based commentaries, and studies lacking methodological transparency. Preference was given to highly cited foundational works and recent contributions (2015–2024) to capture both theoretical foundations and contemporary advancements.

1.4.3 Time Range and Keywords

The review primarily covers publications from 2010 onward, reflecting the rapid growth of deep learning and XAI research. Search keywords included “Explainable AI,” “trust modeling,” “dynamic trust,” “human-AI trust calibration,” “safety-critical AI,” and “trust-aware systems.” Boolean operators were used to refine search queries and ensure comprehensive coverage of interdisciplinary literature.

2. FOUNDATIONS AND THEORETICAL BACKGROUND

The conceptual foundation of dynamic trust-aware explainability lies at the intersection of Explainable Artificial Intelligence (XAI), computational trust modeling, and risk-sensitive system design. Understanding these theoretical pillars is essential for critically analyzing integrated frameworks intended for high-consequence domains. This section synthesizes foundational constructs, formal definitions, and evaluation paradigms that underpin current research.

2.1 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence refers to methods and techniques that make the decision-making processes of AI

systems understandable to humans. The field emerged in response to the opacity of complex machine learning models, particularly deep neural networks, whose internal representations are often non-intuitive. XAI seeks to bridge this interpretability gap by providing human-interpretable reasoning, feature relevance insights, and decision rationales (Gunning, 2017). Conceptually, explainability overlaps with interpretability, transparency, and accountability, though these terms are not strictly synonymous (Doshi-Velez and Kim, 2017).

complicates cross-study comparison and underscores the need for domain-sensitive evaluation frameworks.

2.2 Trust in AI Systems

Trust in AI systems is a socio-technical construct involving psychological, computational, and relational dimensions. It determines the extent to which users rely on automated outputs under uncertainty. In human-automation research, trust is often defined as the attitude that an agent will help achieve goals in situations characterized by vulnerability (Lee and See, 2004). In AI contexts, trust becomes particularly critical when system outputs influence high-stakes decisions.

2.2.1 Cognitive vs Computational Trust

Cognitive trust refers to the human psychological state of reliance based on perceived competence, predictability, and integrity of the system. It is shaped by user experience, explanation clarity, and prior outcomes. Computational trust, in contrast, is mathematically modeled within the system to quantify reliability using probabilistic reasoning, reputation mechanisms, or performance histories. Bayesian trust models, for example, dynamically update reliability estimates as new evidence becomes available (Yu, Singh and Sycara, 2004). The distinction between cognitive and computational trust highlights the need to align system-level trust metrics with human perception to avoid trust miscalibration.

2.2.2 Human-AI Trust Calibration

Trust calibration refers to aligning user reliance with actual system capability. Over-trust can result in automation bias, whereas under-trust leads to disuse or algorithm aversion (Dietvorst, Simmons and Massey, 2015). Effective calibration requires transparent feedback mechanisms, uncertainty quantification, and context-aware explanations. Adaptive explanation strategies that adjust depth or modality according to user expertise have been proposed as mechanisms for improving calibrated reliance in complex environments.

2.3 Characteristics of High-Consequence Decision Domains

High-consequence decision domains are environments where AI errors may result in severe physical, ethical, financial, or societal harm. These domains exhibit high uncertainty, strict accountability requirements, and limited tolerance for failure. Examples include medical diagnosis, aviation control, military operations, and financial risk modeling.

2.3.1 Risk, Uncertainty, and Ethical Impact

Such domains are characterized by probabilistic uncertainty, incomplete information, and potentially irreversible

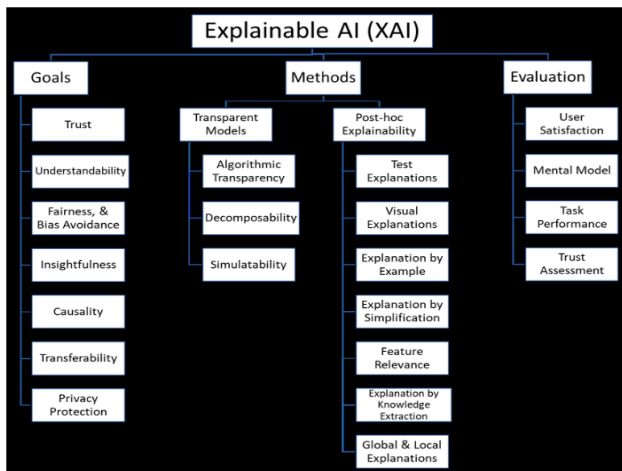


Figure-1: XAI Design and Evaluation Framework

2.1.1 Definitions and Taxonomy: Post-hoc vs Intrinsic Explanations

XAI methods are generally categorized into intrinsic (ante-hoc) and post-hoc approaches. Intrinsic explainability refers to models that are inherently interpretable by design, such as decision trees, rule-based classifiers, and linear models. These models prioritize structural transparency but may sacrifice predictive performance in complex tasks. Post-hoc explainability, by contrast, is applied after model training to interpret otherwise opaque systems. Techniques such as LIME and SHAP generate local feature attributions to approximate decision behavior without modifying the original model architecture (Ribeiro, Singh and Guestrin, 2016; Lundberg and Lee, 2017). While post-hoc approaches are flexible and widely applicable, concerns remain regarding fidelity and stability of generated explanations.

2.1.2 Evaluation Metrics of Explainability

Evaluating explainability remains an open research challenge. Metrics are commonly divided into fidelity (how accurately explanations reflect the underlying model), interpretability (human comprehensibility), and usefulness (impact on decision quality). Quantitative evaluation often measures faithfulness or consistency under perturbation, whereas human-centered evaluation assesses trust, understanding, and cognitive workload (Doshi-Velez and Kim, 2017). The absence of standardized benchmarks

outcomes. Ethical considerations—such as fairness, transparency, and responsibility—become central design constraints. Algorithmic bias or opaque decision-making in healthcare or finance can exacerbate social inequities (Barocas and Selbst, 2016). Consequently, AI systems must not only achieve technical robustness but also provide ethically interpretable reasoning processes.

2.3.2 Regulatory Requirements and Explain ability Standards

Regulatory frameworks increasingly mandate transparency and human oversight for high-risk AI applications. The European Commission’s guidelines for trustworthy AI emphasize requirements including transparency, accountability, and technical robustness (European Commission, 2019). Emerging standards advocate documentation practices, auditability, and explain ability mechanisms that enable traceability of automated decisions. Compliance pressures therefore act as a driving force behind the development of dynamic trust-aware explain ability models capable of satisfying both operational and legal expectations.

3. LITERATURE REVIEW

This section critically synthesizes existing scholarship on explainability models, trust mechanisms, and their integration into trust-aware explainable AI frameworks. Rather than treating explainability and trust as independent constructs, recent research increasingly views them as interdependent components of reliable AI systems, particularly in high-consequence domains.

3.1 Explainability Models in AI

Explainability models aim to make AI decision processes interpretable to stakeholders with varying levels of technical expertise. The literature broadly distinguishes between intrinsically interpretable models and post-hoc explanation techniques designed for complex black-box systems.

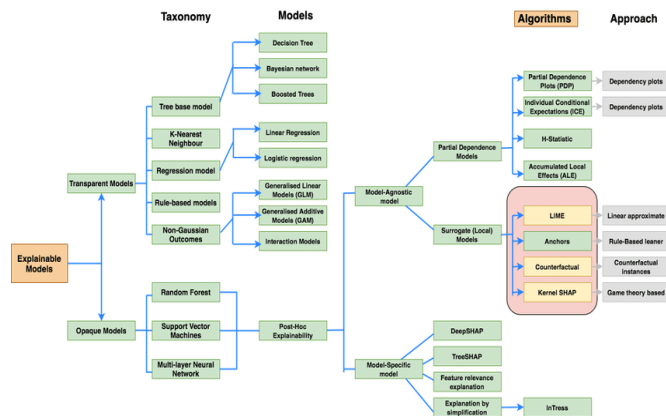


Figure-2: Explainable AI Taxonomy

3.1.1 Intrinsic Explain ability

Intrinsic explain ability refers to models that are transparent by construction. Decision trees, rule-based classifiers, sparse linear models, and generalized additive models are frequently cited as inherently interpretable due to their explicit structural logic. For instance, decision trees provide hierarchical rule paths that can be directly inspected and validated by domain experts (Breiman et al., 1984). Similarly, rule-based systems articulate human-readable IF-THEN logic, facilitating traceability in regulated sectors.

However, intrinsic models often face scalability and expressiveness limitations in high-dimensional, non-linear environments. As data complexity increases—such as in medical imaging or real-time perception tasks—simpler interpretable models may underperform compared to deep neural architectures. This performance-interpretability trade-off remains a central tension in XAI research (Rudin, 2019).

3.1.2 Post-hoc Explain ability

Post-hoc explain ability methods are applied after model training to interpret complex black-box systems. Local surrogate approaches, such as LIME, approximate decision boundaries around specific instances to generate feature importance explanations (Ribeiro, Singh and Guestrin, 2016). SHAP employs cooperative game theory to compute Shapley values, attributing contributions of individual features to predictions (Lundberg and Lee, 2017).

Attention-based explanations in deep learning further aim to highlight salient input regions influencing outputs, particularly in natural language processing and computer vision. Complementary visualization tools—saliency maps, partial dependence plots, and interactive dashboards—enhance user interpretability by presenting explanations graphically. Nevertheless, post-hoc methods raise concerns regarding explanation fidelity and robustness under perturbations.

3.1.3 Comparative Analysis

Intrinsic approaches provide high structural transparency and regulatory alignment but may lack predictive power in complex domains. Post-hoc techniques offer flexibility and compatibility with high-performance models but may introduce approximation errors or misleading rationalizations. Empirical comparisons suggest that while post-hoc explanations improve user understanding, they do not automatically guarantee calibrated trust (Jacovi and Goldberg, 2020). Consequently, the literature indicates a need for hybrid strategies that balance interpretability with accuracy and contextual reliability.

3.2 Trust Mechanisms in AI

Trust mechanisms determine how system reliability is quantified, communicated, and adapted over time. The literature distinguishes between static trust metrics and adaptive, dynamic trust modeling approaches.

3.2.1 Static Trust Models

Static trust models typically rely on predefined rules, performance statistics, or confidence scores to estimate system reliability. Confidence calibration techniques, such as probability scaling and reliability diagrams, provide a numerical representation of prediction certainty. These approaches are common in classification systems where prediction confidence is used as a proxy for trustworthiness.

However, static trust metrics assume environmental stability and consistent data distributions. In safety-critical contexts, such assumptions rarely hold due to concept drift and adversarial conditions. As a result, static models may misrepresent actual reliability under changing operational scenarios (Guo et al., 2017).

3.2.2 Dynamic Trust Models

Dynamic trust models update reliability estimates based on contextual feedback, environmental shifts, and user interaction history. Bayesian trust modeling frameworks adjust posterior trust probabilities as new evidence becomes available, allowing systems to account for uncertainty and performance degradation (Yu, Singh and Sycara, 2004).

Reinforcement learning-based trust adaptation and performance monitoring mechanisms have also been proposed to address model drift in real-time environments. Such adaptive approaches are particularly relevant in autonomous systems and mission-critical infrastructures, where rapid recalibration is necessary to maintain safe human-AI collaboration.

3.3 Trust-Aware Explainability Models

Emerging research integrates trust estimation with explainability mechanisms to create context-aware AI systems capable of supporting calibrated human reliance.

3.3.1 Integrated Frameworks for Trust and Explainability

Integrated frameworks combine model explanations with dynamic trust metrics to modulate the depth, format, or frequency of explanations. For example, adaptive explanation systems tailor outputs based on estimated user expertise and task risk level. Some architectures incorporate uncertainty quantification alongside feature attribution, enabling users to assess both reasoning and reliability simultaneously (Sokol and Flach, 2020).

3.3.2 Domain-Specific Applications

In healthcare diagnostics, explainability tools are used to justify model predictions in radiology and pathology, improving clinician oversight while supporting accountability (Esteve et al., 2017). In autonomous driving, explanation systems clarify perception outputs and decision policies to enhance driver trust and situational awareness.

3.3.3 Evaluation Metrics and Benchmarks

Evaluation of trust-aware explainability models extends beyond predictive accuracy. Human judgment alignment measures assess whether explanations improve user understanding and decision quality. Objective trust quantification involves metrics such as calibration error, reliability curves, and dynamic performance tracking. Despite growing interest, standardized benchmarks integrating both trust calibration and explanation fidelity remain limited, complicating cross-domain validation efforts.

3.4 Comparative and Critical Synthesis

A comparative synthesis of the literature reveals both conceptual convergence and persistent research gaps in integrating trust modeling with explainability techniques.

3.4.1 Comparison of Existing Models

Existing models can be compared across dimensions including explanation type (intrinsic vs post-hoc), trust representation (static vs dynamic), application domain, and evaluation methodology. Studies differ significantly in whether trust is treated as a human perception variable or a computational parameter embedded within the system architecture. Few frameworks comprehensively integrate both dimensions in real-time operational settings.

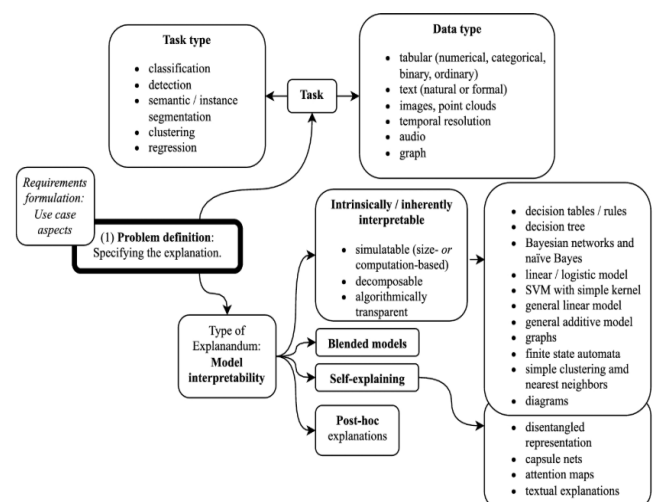


Figure-3: Comprehensive XAI Concept Map

3.4.2 Identified Strengths and Weaknesses

Strengths of current approaches include improved transparency, enhanced uncertainty communication, and better human engagement. However, notable weaknesses persist: limited adaptability under distribution shifts, insufficient modeling of user-specific trust profiles, and inadequate empirical validation in high-risk real-world deployments. Furthermore, many systems focus on explanation generation without empirically measuring long-term trust calibration effects.

3.4.3 Trends in Trust-Aware Explainability Research

Recent trends emphasize interactive explanation interfaces, incorporation of human feedback loops, and regulatory-driven transparency requirements. The increasing prominence of AI governance frameworks has accelerated research on auditable and accountable AI systems. Trust-aware explainability is progressively viewed not only as a usability enhancement but also as a compliance mechanism aligning AI deployment with emerging legal and ethical standards.

4. CHALLENGES AND OPEN ISSUES

Despite substantial progress in explainable AI and computational trust modeling, several unresolved challenges limit the effective deployment of dynamic trust-aware explainability models in high-consequence domains. These challenges span technical scalability, epistemic evaluation, ethical compliance, and human-machine interaction complexities. Addressing them is essential for advancing robust, context-sensitive AI systems.

4.1 Scalability and Real-Time Constraints

One of the foremost challenges concerns scalability in large-scale, high-dimensional systems. Many post-hoc explanation techniques require repeated model evaluations or surrogate approximations, which can introduce computational overhead incompatible with real-time decision environments. For example, model-agnostic explanation methods may not meet latency constraints in autonomous navigation or critical care monitoring systems where milliseconds matter (Adadi and Berrada, 2018).

Furthermore, dynamic trust modeling necessitates continuous monitoring of system performance, uncertainty estimation, and context-aware recalibration. Integrating these processes into resource-constrained edge or embedded systems amplifies architectural complexity. Ensuring low-latency explanation generation without sacrificing fidelity remains an open systems engineering problem.

4.2 Balancing Performance vs Interpretability

The performance-interpretability trade-off continues to pose theoretical and practical tensions. Highly expressive deep learning architectures often outperform interpretable models in complex tasks such as image recognition or anomaly detection. However, their opacity complicates accountability and human oversight (Rudin, 2019).

Attempts to embed interpretability directly within neural networks—such as attention mechanisms or inherently interpretable neural modules—have shown promise but may still lack rigorous guarantees of transparency. Achieving both predictive optimality and robust interpretability without resorting to approximate post-hoc explanations remains a central research objective in safety-critical AI design.

4.3 Measuring Trust and Explanation Quality

A significant methodological challenge lies in quantifying trust and explanation effectiveness in a standardized manner. Trust is inherently psychological and context-dependent, making it difficult to operationalize as a stable metric. While calibration error and uncertainty measures capture statistical reliability, they do not fully represent user perception or behavioral reliance (Lee and See, 2004).

Similarly, explanation quality lacks universal benchmarks. Fidelity, stability, completeness, and human comprehensibility are often evaluated separately, leading to fragmented assessment frameworks. The absence of domain-specific evaluation standards for integrated trust-explainability systems impedes reproducibility and cross-study comparability.

4.4 Ethical and Legal Considerations

Ethical and legal constraints significantly shape the deployment of AI systems in high-consequence settings. Regulatory frameworks increasingly mandate transparency, fairness, and accountability in automated decision-making. For instance, European AI governance guidelines emphasize explainability and human oversight as prerequisites for high-risk systems (European Commission, 2019).

However, explainability alone does not guarantee fairness or absence of bias. Algorithmic decisions may embed structural inequities that persist despite transparent reasoning (Barocas and Selbst, 2016). Additionally, disclosing detailed explanations may conflict with intellectual property protection or security considerations. Balancing transparency with privacy, safety, and proprietary constraints presents a multidimensional governance challenge.

4.5 Human–Machine Teaming and Cognitive Load

Effective trust-aware explainability must account for human cognitive limitations. Overly complex or frequent explanations can increase cognitive load, impair situational awareness, and reduce decision efficiency. Research in human–automation interaction indicates that optimal transparency is context-dependent; excessive detail may overwhelm users, whereas insufficient explanation undermines trust calibration (Parasuraman, Sheridan and Wickens, 2000).

Dynamic explanation systems must therefore adapt to user expertise, stress levels, and task criticality. Designing interfaces that support collaborative human–AI teaming—without inducing automation complacency or overload—requires interdisciplinary integration of cognitive science, human–computer interaction, and AI system design principles.

5. CONCLUSION

This review has critically examined the evolution of explainable artificial intelligence and computational trust modeling, with particular emphasis on their integration into dynamic trust-aware explainability frameworks for high-consequence decision domains. The analysis demonstrates that while significant advancements have been achieved in post-hoc explanation techniques, intrinsic interpretability methods, and probabilistic trust modeling, their integration remains fragmented and context-insensitive. Existing systems frequently address explainability and trust as parallel objectives rather than interdependent design requirements. In safety-critical environments such as healthcare, autonomous systems, finance, and defense, static explanation mechanisms and fixed trust metrics are insufficient to ensure calibrated human reliance under uncertainty and distributional shifts. The literature highlights the importance of adaptive explanation strategies, uncertainty quantification, and user-centered evaluation for enhancing accountability and collaborative decision-making. However, standardized benchmarks for jointly evaluating trust calibration and explanation fidelity are still lacking. Future progress requires interdisciplinary approaches that integrate human factors engineering, regulatory compliance, and real-time system optimization. Ultimately, dynamic trust-aware explainability models represent a foundational step toward responsible, transparent, and resilient AI systems capable of operating reliably in environments where decisions carry significant ethical, societal, and operational consequences.

5.1. Limitations of the Review

This review is subject to several limitations. First, although major academic databases were consulted, the synthesis primarily emphasizes peer-reviewed literature in English, potentially excluding relevant contributions from non-

indexed or regional publications. Second, the rapidly evolving nature of explainable AI and trust modeling means that newly emerging frameworks may not be fully represented. Third, empirical comparisons across studies were constrained by heterogeneous evaluation metrics and domain-specific methodologies, limiting the ability to conduct quantitative cross-study analysis. Additionally, the review focuses predominantly on technical and conceptual frameworks rather than conducting experimental validation or meta-analytic statistical synthesis. Finally, regulatory discussions are contextualized primarily within widely cited global guidelines and may not comprehensively reflect jurisdiction-specific legal nuances.

REFERENCES

1. Adadi, A. and Berrada, M. (2018) 'Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)', *IEEE Access*, 6, pp. 52138–52160.
2. Barocas, S. and Selbst, A.D. (2016) 'Big data's disparate impact', *California Law Review*, 104(3), pp. 671–732.
3. Bonnefon, J.F., Shariff, A. and Rahwan, I. (2016) 'The social dilemma of autonomous vehicles', *Science*, 352(6293), pp. 1573–1576.
4. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Belmont, CA: Wadsworth.
5. Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015) 'Algorithm aversion: People erroneously avoid algorithms after seeing them err', *Journal of Experimental Psychology: General*, 144(1), pp. 114–126.
6. Doshi-Velez, F. and Kim, B. (2017) 'Towards a rigorous science of interpretable machine learning', *arXiv preprint arXiv:1702.08608*.
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017) 'Dermatologist-level classification of skin cancer with deep neural networks', *Nature*, 542(7639), pp. 115–118.
8. European Commission (2019) *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission High-Level Expert Group on Artificial Intelligence.
9. Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) 'On calibration of modern neural networks', in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330.
10. Gunning, D. (2017) 'Explainable Artificial Intelligence (XAI)', *Defense Advanced Research Projects Agency (DARPA)*. Available at:

- <https://www.darpa.mil/program/explainable-artificial-intelligence>.
11. Jacovi, A. and Goldberg, Y. (2020) 'Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?', in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4198–4205.
 12. Lee, J.D. and See, K.A. (2004) 'Trust in automation: Designing for appropriate reliance', *Human Factors*, 46(1), pp. 50–80.
 13. Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', in *Advances in Neural Information Processing Systems (NeurIPS)*, 30, pp. 4765–4774.
 14. Parasuraman, R., Sheridan, T.B. and Wickens, C.D. (2000) 'A model for types and levels of human interaction with automation', *IEEE Transactions on Systems, Man, and Cybernetics – Part A*, 30(3), pp. 286–297.
 15. Ribeiro, M.T., Singh, S. and 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 (KDD), pp. 1135–1144.
 16. Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, 1(5), pp. 206–215.
 17. Sokol, K. and Flach, P. (2020) 'One explanation does not fit all: The promise of interactive explanations for machine learning transparency', *KI – Künstliche Intelligenz*, 34, pp. 235–250.
 18. Yu, B., Singh, M.P. and Sycara, K. (2004) 'Developing trust in large-scale peer-to-peer systems', in Proceedings of the IEEE First Symposium on Multi-Agent Security and Survivability, pp. 1–10.
 19. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K. and Teevan, J. (2019) 'Guidelines for human-AI interaction', Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13.
 20. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R. and Herrera, F. (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58, pp. 82–115.
 21. Bansal, G., Nushi, B., Kamar, E., Horvitz, E., Weld, D.S. and Lasecki, W.S. (2019) 'Beyond accuracy: The role of mental models in human-AI team performance', Proceedings of AAAI Conference on Artificial Intelligence, 33(01), pp. 9669–9676.
 22. Bhatt, U., Andrus, M., Weller, A. and Xiang, A. (2020) 'Machine learning explainability for external stakeholders', Proceedings of FAT Conference*, pp. 344–353.
 23. Carvalho, D.V., Pereira, E.M. and Cardoso, J.S. (2019) 'Machine learning interpretability: A survey on methods and metrics', *Electronics*, 8(8), 832.
 24. Chen, J.Y.C., Barnes, M.J. and Harper-Sciarini, M. (2011) 'Supervisory control of multiple robots: Human-performance issues and user-interface design', *IEEE Transactions on Systems, Man, and Cybernetics*, 41(2), pp. 435–454.
 25. Endsley, M.R. (2017) 'From here to autonomy: Lessons learned from human-automation research', *Human Factors*, 59(1), pp. 5–27.
 26. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. and Kagal, L. (2018) 'Explaining explanations: An overview of interpretability of machine learning', IEEE 5th International Conference on Data Science and Advanced Analytics, pp. 80–89.
 27. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., De Visser, E.J. and Parasuraman, R. (2011) 'A meta-analysis of factors affecting trust in human-robot interaction', *Human Factors*, 53(5), pp. 517–527.
 28. Holzinger, A., Carrington, A. and Müller, H. (2020) 'Measuring the quality of explanations: The system causability scale', *KI – Künstliche Intelligenz*, 34, pp. 193–198.
 29. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H. and Wortman Vaughan, J. (2020) 'Interpreting interpretability: Understanding data scientists' use of interpretability tools', Proceedings of CHI Conference on Human Factors in Computing Systems, pp. 1–14.
 30. Lipton, Z.C. (2018) 'The mythos of model interpretability', *Queue*, 16(3), pp. 31–57.
 31. Miller, T. (2019) 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, 267, pp. 1–38.
 32. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B. and Snoek, J. (2019) 'Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift',

Advances in Neural Information Processing Systems,
32.

33. Wachter, S., Mittelstadt, B. and Russell, C. (2017) 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR', *Harvard Journal of Law & Technology*, 31(2), pp. 841–887.
34. Zhang, Y. and Dafoe, A. (2019) 'Artificial intelligence: American attitudes and trends', Center for the Governance of AI Working Paper, University of Oxford.