How Can Artificial Intelligence Explain its Thinking? A Review of Explainable AI

Naren Chandra
Acton, MA, USA

naren1712@gmail.com

Abstract: Artificial Intelligence (AI) systems are increasingly shaping decision-making across industries such as healthcare, finance, e-commerce, education, and cybersecurity. Despite significant advances in machine learning and generative AI, many AI models continue to function as "black boxes," producing outputs and recommendations without clearly communicating the reasoning behind them. This lack of transparency has raised growing concerns regarding trust, accountability, fairness, and ethical AI adoption. As organizations become more dependent on AI-driven systems, explainability has emerged as a critical factor influencing how users perceive, trust, and adopt artificial intelligence technologies. This literature review examines how Explainable Artificial Intelligence (XAI) enables AI systems to communicate their reasoning in ways understandable to human users. The paper analyzes existing academic and industry research on explainability, transparency, and interpretability within AI systems, with a particular focus on how explanations influence human trust and organizational acceptance of AI technologies. The review evaluates major explainability approaches used in both traditional machine learning models and modern generative AI systems, including techniques such as Local Interpretable Model-Agnostic Explanations (LIME), SHapley Additive exPlanations (SHAP), attention visualization, counterfactual explanations, and chain-of-thought reasoning methods used in large language models (LLMs). In addition, the paper examines the distinctions between transparency, interpretability, and explainability, and how these concepts contribute to building trustworthy AI systems. The review further analyzes explainability from a human-centered perspective, including perceived fairness, ethical AI design, user confidence, and organizational adoption challenges. The paper also discusses the growing importance of explainability within responsible AI governance frameworks, particularly as regulatory initiatives such as the European Union AI Act increasingly emphasize transparency and accountable AI decision-making. This study contributes to the growing body of literature on explainable AI by consolidating fragmented research on transparency, trust, and human-centered AI adoption. The review identifies current research gaps, limitations of existing explainability approaches, and future opportunities for designing AI systems that are not only accurate and efficient but also understandable, trustworthy, and ethically aligned with human values.

Keywords: Explainable artificial intelligence (XAI), Artificial intelligence transparency, Human trust in AI, Interpretable machine learning, Responsible AI governance

1. Introduction

Artificial intelligence (AI) has moved from being a tool for specialized analytical use to that of being a general-purpose infrastructure for making predictions, recommendations, generating content and supporting decision making. Machine learning systems now influence clinical diagnosis, credit assessment, education, e-commerce, cybersecurity, and public services. Generative AI extends this reach by producing language, images, code, and proposed actions through large models that are often accessed by non-technical users. The organizational value is huge, but so is the information asymmetry between a system that produces an answer and a user who must decide whether the answer provided is reliable.

The primary issue or challenge for a user is that of opacity. Deep neural networks, and large language models can contain many interacting parameters whose internal operations may not be easily decipherable to human-readable rules. A model may be accurate overall, while not being able to provide an explanation of why a particular applicant was denied credit, why a patient was classified as high risk, or why a security alert was generated. Reviews of black-box models identify this gap as a barrier to adoption in high-stakes settings because affected people, professionals, auditors, and regulators need different forms of justification (Hassija et al., 2024; Linardatos et al., 2021; Mittelstadt, 2023; Chandra, 2025).

Explainable AI (XAI) addresses this gap by designing models that are opaque by adding explanation methods that make them easy to understand. Explanations help users judge when to accept or override an AI recommendation. They help reveal correlations, support error correction, communicate uncertainty and make decisions contestable. However, explanation is not synonymous with trustworthiness as a persuasive explanation with a flawed logic may increase confidence without improving reliability. Trust must be calibrated to demonstrate system competence, enhance explanation quality, and explain the consequences of error (Miller, 2019; Nalepa et al., 2019).

This review synthesizes literature across the domains of computer science, information systems, human-computer interaction and governance. Its objective is to explain how major XAI techniques work and to evaluate

how explanation affects human trust, organizational adoption, and responsible oversight. The analysis is organized around two research questions:

RQ1: How do AI systems explain their decisions and recommendations?

RQ2: how does explainability influence human trust and adoption of AI systems?

2. Explainable AI: Concepts and Foundation

XAI is an umbrella term for methods that make an AI system's behavior, outputs, or governing process understandable to a relevant human audience. The definition is audience and purpose dependent. A developer may need feature-level diagnostics, a clinician may need evidence linked to a diagnosis, a consumer may need a concise reason and a route to appeal, and an auditor may need documentation, logs, and reproducible tests. Accordingly, an explanation should specify what is being explained, to whom, for what decision, and with what evidence (Chinnaraju, 2025; Miller, 2019; Mittelstadt, 2023; Thalpage, 2023).

Three related concepts are often equated in contemporary literature. Transparency concerns what information is disclosed about a system, including its data, purpose, model, limitations, and governance. Interpretability concerns the degree to which a person can understand the relationship between inputs, internal structure, and outputs. Explainability concerns a communicative account of why or how an output occurred. An interpretable decision tree may expose its complete path, whereas a black-box model may remain structurally opaque but provide a local SHAP (Shapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) explanation. Transparency can therefore support explainability without guaranteeing that an explanation is accurate, complete, or useful (Balasubramaniam et al., 2022; Lipton, 2018).

Explainability matters because accuracy alone cannot answer questions of fairness, contestability, safety, and responsibility. Explanations allow stakeholders to inspect whether a decision relied on relevant factors, identify bias or data leakage, and determine whether a system is being used within its intended scope. In cybersecurity and medicine, recent applications show that feature-attribution models can make high-performing models more actionable for analysts and clinicians, although practical value depends on stability and domain validation (Agarwal et al., 2025; Mohale & Obagbuwa, 2025a).

The evolution of XAI reflects shifts in AI itself. Early symbolic and rule-based systems exposed explicit knowledge and inference rules. Statistical learning introduced relatively interpretable coefficients and trees. As ensemble and deep-learning models improved predictive performance, post-hoc methods such as LIME and SHAP became prominent. Generative AI has now moved the debate from feature attribution toward natural-language rationales, evidence traces, and the faithfulness of chain-of-thought. This trajectory changes the question from whether a system can produce an explanation to whether the explanation is causally connected to the output and useful for a particular human decision (Arrieta et al., 2020; Mathew et al., 2025).

Table 1: Distinguishing Core XAI Concepts

Concept	Primary question	Typical evidence	Key limitation
Transparency	What is disclosed about the system and its governance?	Data provenance, documentation, intended use, model cards, logs	Disclosure can be extensive yet still incomprehensible or incomplete.
Interpretability	Can a person understand the model or input-output relationship?	Coefficients, rules, decision paths, monotonic effects	Understanding depends on model complexity, domain knowledge, and audience.
Explainability	Why or how did this output occur?	Attributions, examples, counterfactuals, rationales	A plausible explanation may be unstable or unfaithful to the model.
Accountability	Who can evaluate, challenge, and remedy the decision?	Audit trails, oversight roles, impact assessments, appeals	Explanation without authority or a remedy may not change outcomes.

Note. Synthesized from Balasubramaniam et al. (2022), Lipton (2018), Mittelstadt (2023), and Tabassi (2023).

3. Major Explainability Techniques

3.1 Intrinsic Explainability

Intrinsically explainable models make their relevant logic available through the model structure itself. Decision trees partition the feature space through a sequence of tests: a prediction can be traced along a path from root to leaf. They are intuitive and can expose interactions, but deep or heavily pruned trees may become unstable or difficult to comprehend. Rule-based systems represent knowledge as explicit if-then conditions. They are inspectable and can encode domain constraints, although large rule sets can conflict, become brittle, fail to capture subtle nonlinear patterns. Linear and logistic regression express predictions as weighted combinations of features. Their coefficients can communicate direction and magnitude when variables are well specified, but correlation, interactions, transformations, and scale complicate causal interpretation (Linardatos et al., 2021).

Intrinsic models are particularly attractive in high-stakes settings because the same structure performs the prediction and supports the explanation. Rudin (2019) argues that interpretable models should be preferred over black-box explanations when a comparably accurate transparent model is feasible. The trade-off is contextual rather than universal: a simple model can be more trustworthy when the data-generating process is compatible with its assumptions, but simplicity alone does not ensure fairness, correctness, or adequate performance.

3.2 Post-hoc Explainability

LIME

Local Interpretable Model-Agnostic Explanations (LIME) explains one prediction by modifying the input near the instance of interest, observing the black-box model's responses, and fitting a simple weighted surrogate model in that local neighborhood. The resulting feature weights indicate which inputs supported or opposed the prediction. LIME is a flexible across model classes and useful for local diagnosis and user-facing explanations. Its limitations arise from the same approximation process – explanations can change with the modification distribution, random seed, feature representation, or definition of locality, and the local surrogate may not describe global behavior (Ribeiro et al., 2016).

SHAP

Shapley Additive exPlanations (SHAP) assigns each feature a contribution value derived from cooperative game theory. A prediction is compared with a baseline, and the difference is allocated across features according to their marginal contributions over possible feature coalitions. SHAP provides a unified additive framework with desirable consistency and local-accuracy properties, and aggregated values can support both local and global analysis. However, exact computation is expensive, approximations may be necessary, and results depend on the baseline and assumptions about feature dependence. A SHAP value describes contribution relative to a model and background distribution; it does not establish causality (Lundberg & Lee, 2017).

Counterfactual Explanations

Counterfactual explanations identify a minimally different input that would have changed the model's outcome: for example, the income or debt ratio at which a rejected loan application would have been approved. They are contrastive, actionable, and often intuitive because they answer, "What would need to be different?" They can also support contestability without disclosing the full model. Their quality depends on feasibility, proximity, sparsity, and causal realism. A mathematically small change may be impossible, ethically inappropriate, or outside the person's control, and multiple valid counterfactuals may convey different policy implications (Wachter et al., 2018).

Attention Visualization

Attention visualizations display the weights a neural model assigns to tokens, image regions, or other inputs during computation. They can help users explore associations and may reveal where a model concentrates information. Yet attention weight is not automatically an explanation of causal influence. Jain and Wallace (2019) showed that alternative attention patterns can yield similar predictions and that attention may correlate poorly with other importance measures; Wiegreffe and Pintner (2019) responded that attention can still have qualified explanatory value when the explanatory claim and evaluation are carefully specified. Attention is therefore best treated as diagnostic evidence, not a complete account of reasoning.

Chain-of-Thought Reasoning in Large Language Models

Chain-of-thought (CoT) prompting elicits intermediate natural-language steps before a large language model gives an answer. CoT can improve performance on multi-step tasks and makes a response easier for users to follow (Wei et al., 2022). However, a generated rationale is an output, not privileged access to hidden internal computation. Models may omit influential cures, rationalize a biased answer after the fact, or produce a fluent explanation that remains factually wrong. Intervention studies demonstrate that stated reasons can be unfaithful to the actual determinants of a response (Turpin et l., 20243)). For governance and high stakes use, concise rationales should therefore be paired with verifiable evidence, uncertainty, tool or retrieval traces, and tests showing that changing a stated reason changes the output in the expected direction.

3.3 Explainability for Foundation Models and Large Language Models

Foundation models complicate explainability because one model performs many tasks and its outputs vary with prompts, context, sampling and updates. Recent approaches therefore combine generated rationales with retrieval citations, too-use logs, provenance, and causal interventions that test whether stated reasons influence the answer. Mechanistic interpretability examines activations and circuits; sparse autoencoders can identify more interpretable internal features, although completeness and causal validation remain unresolved (Cunningham et al., 2024). These approaches introduce substantial computational and governance trade-offs. Repeated sampling, counterfactual prompting, activation capture, and circuit analysis can be expensive at foundation-model scale, while disclosing detailed reasoning or traces may expose private data, security sensitive information, or opportunities for system gaming. For foundation models, natural language rationales should be combined with retrieval or tool provenance and causal faithfulness tests rather than treated as self-authenticating evidence(Agrawal et al., 2025; Mohale & Obagbuwa, 2025a; Rane et al., 2023). Table 2 summarizes these trade-offs.

3.4 Critical Comparison and Domain Suitability

No technique is uniformly superior. Intrinsic models offer low-cost global inspection but may miss complex patterns. LIME is flexible but unstable; SHAP is principled but computationally demanding and counterfactuals are actionable only when feasible. In healthcare, SHAP or interpretable clinical rules can support professional review, but explanations must be paired with uncertainty and domain validation. In cybersecurity, latency and adversarial manipulation favor efficient local diagnostics and robust aggregation rather than expensive instance-by-instance optimization.

3.5 Evaluating Explanation Quality

Explanation quality is multi-dimensional. Fidelity measures agreement with model behavior, while faithfulness asks whether the stated factors causally influence the output. Stability examines consistency across repeated runs or similar inputs, and robustness tests sensitivity to noise, manipulation, and distribution shifts. Human-centered evaluation should additionally measure comprehension, task performance, error detection, decision times, reliance calibration, and fairness. Since these criteria can conflict, evaluation should combine functional tests, robustness checks, and studies with the intended users (Doshi-Velez & Kim, 2017: Miller, 2019).

Table 2: Critical Comparison of Major Explainability Techniques

Technique	Interpretability and scope	Model compatibility	Relative cost	Practical applications	Main trade-offs
Decision trees	High; global and local decision paths	Primarily tabular supervised models	Low	Clinical rules, eligibility and risk scoring	Deep trees become unstable and may underfit complex patterns.
Rules / linear models	High when compact; mainly global	Tabular and additive relationships	Low	Credit, policy and compliance models	May miss nonlinear interactions; coefficients are not causal effects.
LIME	Moderate; local approximate surrogate	Model-agnostic; tabular, text and images	Medium	Local debugging and case-level explanations	Sensitive to locality, sampling, representation and randomness.

Technique	Interpretability and scope	Model compatibility	Relative cost	Practical applications	Main trade-offs
SHAP	Moderate to high; local and aggregate attribution	Model-agnostic with optimized variants	High	Model audit, healthcare, finance and feature analysis	Baseline and dependence assumptions affect results; noncausal.
Counterfactuals	High for local contrastive questions	Models with editable inputs	Medium-high	Recourse in lending, employment and insurance	Solutions may be infeasible, unstable, discriminatory or unrealistic.
Attention	Moderate as a local diagnostic	Attention-based neural models	Low-medium	Language, vision and multimodal exploration	Attention weight is not necessarily causal importance.
CoT / rationales	High readability; local language account	Generative language and multimodal models	Medium-high	Tutoring, assisted analysis and decision support	May be persuasive but post-hoc, incomplete or unfaithful.
Mechanistic features	Potentially high internal insight	Foundation models with activation access	Very high	Developer debugging, safety analysis and audit	Requires access and compute; completeness and causal validity remain uncertain.

Note. Technique selection should match the explanatory claim, stakeholder, risk, and evaluation method. LIME = Local

Interpretable Model-Agnostic Explanations; SHAP = SHapley Additive exPlanations.

4. Explainability and Human Trust

From an information systems perspective, the value of XAI is realized through human judgement and organizational use rather than through technical transparency alone. Trust in AI is a willingness to rely on a system under uncertainty; appropriate trust occurs when exists when user reliance directly aligns with the system's actual capabilities. Explanations can improve that calibration by showing relevant evidence, limitations, and conditions under which a recommendation should be questioned. Organizational guidelines reviewed by Balasubramaniam et al. (2022) identify customers and users as central explanation audiences while also recognizing the needs of developers, partners, and other stakeholders.

Meaningful explanations can increase perceived fairness because people can inspect whether a decision used appropriate factors and whether comparable cases appear to be trusted consistently. They can support user acceptance by reducing uncertainty, improve confidence in recommendations by exposing supporting evidence, and strengthen human-AI collaboration by making disagreement productive (Barnes & Hutson, 2024). In clinical and cybersecurity applications, interpretable feature contributions help professionals validate predictions and prioritize investigation rather than merely accept a score (Agrawal et al., 2025; Mohale & Obagbuwa, 2025a). In this sense, XAI can improve adoption, decision support, model debugging, and institutional learning.

The same mechanisms can also produce harm. Overtrust occurs when users mistake fluency, detail, or visual polish for fidelity; undertrust emerges when explanations are too technical, inconsistent, or poorly aligned with professional knowledge. Cognitive overload can arise when a system provides dozens of features or a lengthy rationale without prioritizing the information needed for action. False explanations are especially dangerous because they offer a coherent story that does not accurately represent the causes of the output. Trust building should therefore be iterative and contextual, with explanations tested for comprehension, usefulness, and their effect on reliance over time rather than evaluated only by user satisfaction (Miller, 2019; Nalepa et al., 2019).

A human-centered design strategy separates audiences and decision, Developers may need global error analysis and counterfactual stress tests; domain professionals need evidence and uncertainty; affected individuals need clear reasons, actionable recourse, and accessible language; auditors need documentation and reproducible logs. The goal is not maximum disclosure but sufficient, faithful, and usable information for the task. Explanation quality should be judged by whether it helps users detect errors, calibrate reliance, contest harmful outcomes, and collaborate effectively with the system.

5. Explainability and Responsible AI Governance

Explainability is a governance mechanism because it links technical behavior to accountability. It can make bias visible, support audits, document the basis of decision, and enable human oversight. However, an explanation alone does not assign responsibility or create a remedy. Responsible governance also requires clear ownership, data and model provenance, validation, monitoring, incident reporting, impact assessment, and accessible challenge processes. The NIST AI Risk Management Framework treats transparency, explainability, and interpretability as related but distinct characteristics of trustworthy AI and organizes risk management around govern, map, measure, and manage functions (Tabassi, 2023).

European regulation illustrates why precise legal claims matter. The General Data Protection Regulation (GDPR) does not simply establish an unconditional, universal "right to explanation." Article 22 restricts certain decisions based solely on automated processing that produce legal or similarly significant effects and provides safeguards in specified circumstances, including human intervention and contestation; Articles 13-15 also require meaningful information about the logic involved in relevant automated processing (European Parliament & Council of the European Union, 2016). Ther scope and content of explanation duties remain context-dependent.

The European Union Artificial Intelligence Act establishes more explicit risk-based obligations. Article 13 requires high-risk systems to be sufficiently transparent for deployers to interpret outputs and use them appropriately. Article 50 requires disclosures for certain systems that interact with people or generate synthetic content. Article 86 provides qualifying affected persons and right to clear and meaningful information about the AI system's role and the main elements of certain high-risk decisions. The Act entered into force in 2024 and is scheduled to apply generally from August 2, 2026, subject to phased provisions (European Parliament & Council of the European Union, 2024).

Organizations care because weak explainability can create compliance, litigation, operational, and reputational risk. Conversely, validated explanations can strengthen consumer trust, help identify discriminatory patterns, support internal control, and demonstrate that AI deployment aligns with stated values. The practical implication is to treat explanation requirements as part of system design and procurement, not as a post-deployment interface added after the consequential choices have already been made (Ramachandram et al., 2025).

6. Research Gaps and Future Directions

First, XAI lacks standardized, decision-relevant metrics. Properties like fidelity, stability, completeness, sparsity, fairness, comprehensibility, and usefulness capture vastly different traits, meaning a method can excel at one while failing another. Future evaluation should combine functional tests, causal interventions, robustness checks, and human-subject studies matched to the intended audience and consequence (Arrieta et al., 2020; Doshi-Velez & Kim, 2017; Mathew et al., 2025).

Second, generative AI creates a distinct faithfulness problem. Natural-language rationales can be persuasive while remaining disconnected from model computation or external evidence. Research should compare CoT, retrieval citations, tool traces, counterfactual interventions, and independent verification, with special attention to when exposing internal traces creates privacy, security, or gaming risks. The strongest explanation may be a concise, checkable account linked to sources and uncertainty rather than a long narrative of purported thinking.

Third, causal evidence about trust remains limited. Many studies measure perceived trust or satisfaction immediately after an explanation, but these outcomes do not show whether users make better decisions or remain appropriately calibrated after errors, model updates, and distribution shifts. Longitudinal and field experiments should measure reliance, error detection, override behavior, appeal quality, and downstream fairness. Researchers should also investigate when explanation increases unjustified confidence or shifts responsibility from organizations to users.

Fourth, industry-specific frameworks are needed. A useful explanation for a fraud analyst differs from one for a physician, teacher, regulator, or loan applicant. Sector-specific standards should identify permissible evidence, required uncertainty, time constraints, domain vocabulary, validation procedures, and recourse. Work in healthcare, finance, and intrusion detection demonstrates the value of contextual design while also showing that explanation methods and performance constraints vary across domains (Agrawal et a., 2025; Mohale & Obagbuwa, 2025b; Rane et al., 2023).

Fifth, human-centered XAI requires participatory design. Affected people and frontline professionals should help define what questions explanations must answer, what cognitive load is acceptable, and what actions follow.

Future models should adapt explanation depth to expertise, communicate uncertainty and alternatives, and connect explanations to governance processes. This agenda reframes XAI from "making the model visible" to supporting informed, fair, and contestable human decisions.

7. Conclusion

Explainability is becoming essential to responsible AI adoption because predictive accuracy alone cannot support justified trust, accountability, or contestability. AI systems explain decisions through intrinsically interpretable structures, local and global feature attributions, counterfactuals, visualizations, and natural language rationales, Each method answers a different question and carries distinct assumptions; none provides a universal window into model reasoning.

The evidence indicates that explanations can improve acceptance, perceived fairness, decision support, and human-AI collaboration when they are faithful, understandable, and aligned with stakeholder needs. The same explanations can create overtrust, undertrust, cognitive overload, or false confidence when they are unstable or merely plausible. Future AI systems must therefore balance performance, transparency, and accountability through human-centered evaluation, sector-specific controls, and governance that links explanation to audit, oversight, and remedy. The goal is not AI that narrates every internal operation, but AI whose consequential claims can be understood, tested, and responsibly acted upon.

The next research agenda needs standardized, context sensitive benchmarks combining fidelity, faithfulness, stability, robustness, computational cost, and human decision quality. Foundation model research must determine when rationales reflect causal computation, how mechanistic features scale across models and modalities, and how retrieval and tool traces can be verified without exposing private or security sensitive information. Longitudinal studies should measure whether explanations improve calibrated reliance, error correction, appeals, and fairness after model updates and distribution shifts. Progress requires treating explanation as a socio-technical capability validated for a specific model, audience, domain, and consequence rather than as a universal interface (Cunningham et al., 2024; Jacovi & Goldberg, 2020; Lanham et al., 2023).

Ethics Declaration: Ethical clearance was not required for this study.

AI Declaration: During the preparation of this article, the author used OpenAI to improve the paper's readability. After using this tool, the author reviewed and edited the content as needed.

References

Agrawal, R., Gupta, T., Gupta, S., Chauhan, S., Patel, P., & Hamdare, S. (2025). Fostering trust and interpretability: Integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency. *Diagnostic Pathology, 20*, Article 105. https://doi.org/10.1186/s13000-025-01686-3

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. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Balasubramaniam, N., Kauppinen, M., Hiekkanen, K., & Kujala, S. (2022). Transparency and explainability of AI systems: Ethical guidelines in practice. In V. Gervasi & A. Vogelsang (Eds.), *Requirements engineering: Foundation for software quality* (pp. 3–18). Springer. https://doi.org/10.1007/978-3-030-98464-9_1

Barnes, E., & Hutson, J. (2024). Navigating the complexities of AI: The critical role of interpretability and explainability in ensuring transparency and trust. *International Journal of Multidisciplinary and Current Educational Research, 6(3)*, 248–256.

Chandra, N. (Ed.). (2025). *Organizational and social impact of AI-Driven fake review detection in e-commerce: Vol. 6 No. (1)* (6th International Conference on AI Research) [University of Genoa]. Academic Conferences and Publishing International Ltd. https://papers.academic-conferences.org/index.php/icair/article/view/4207

Chinnaraju, A. (2025). Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. *World Journal of Advanced Engineering Technology and Sciences, 14(3)*, 170–207. https://doi.org/10.30574/wjaets.2025.14.3.0106

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. https://arxiv.org/abs/1702.08608

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union, L 119, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation, 16*, 45–74. https://doi.org/10.1007/s12559-023-10179-8

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 3543–3556). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1357

Jacovi, A., & 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* (pp. 4198-4205). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.386

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., . . . Perez, E. (2023). *Measuring faithfulness in chain-of-thought reasoning*. arXiv. https://arxiv.org/abs/2307.13702

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23(1)*, Article 18. https://doi.org/10.3390/e23010018

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61(10)*, 36–43. https://doi.org/10.1145/3233231

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates.

Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent emerging techniques in explainable artificial intelligence to enhance the interpretable and understanding of AI models for human. *Neural Processing Letters, 57*, Article 16. https://doi.org/10.1007/s11063-025-11732-2

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B. (2023). *Interpretability and transparency in artificial intelligence*. SSRN. https://doi.org/10.2139/ssrn.4317938

Mohale, V. Z., & Obagbuwa, I. C. (2025a). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. *Frontiers in Computer Science, 7*, Article 1520741. https://doi.org/10.3389/fcomp.2025.1520741

Mohale, V. Z., & Obagbuwa, I. C. (2025b). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. *Frontiers in Artificial Intelligence, 8*, Article 1526221. https://doi.org/10.3389/frai.2025.1526221

Nalepa, G. J., Araszkiewicz, M., Nowaczyk, S., & Bobek, S. (2019). Building trust to AI systems through explainability: Technical and legal perspectives. *CEUR Workshop Proceedings, 2681*, 1–6. https://ceur-ws.org/Vol-2681/xaila2019-paper2.pdf

Ramachandram, D., Joshi, H., Zhu, J., Gandhi, D., Hartman, L., & Raval, A. (2025). *Transparent AI: The case for interpretability and explainability*. arXiv. https://arxiv.org/abs/2507.23535

Rane, N. L., Choudhary, S. P., & Rane, J. (2023). *Explainable artificial intelligence (XAI) approaches for transparency and accountability in financial decision-making*. SSRN. https://doi.org/10.2139/ssrn.4640316

Ribeiro, M. T., 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). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

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

Thalpage, N. S. (2023). Unlocking the black box: Explainable artificial intelligence (XAI) for trust and transparency in AI systems. *Journal of Digital Art & Humanities, 4(1)*, 31–36. https://doi.org/10.33847/2712-8148.4.1_4

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems* (Vol. 36). Curran Associates.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31(2)*, 841–887.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 24824–24837). Curran Associates.

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 11–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002