Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Explainability in Artificial Intelligence: Giving a Method to the Madness

Supriya Medapati
Massachusetts Institute of Technology, USA

ARTICLE INFO

Received: 12 July 2025

Revised: 26 Aug 2025

Accepted: 06 Sept 2025

ABSTRACT

Explainable AI (XAI) has emerged as a critical response to the increasing opacity of advanced machine learning systems, particularly as their predictive capabilities grow while comprehensibility diminishes. This comprehensive article examines the evolution of explainability methods across three primary categories: model-agnostic approaches that function independently of underlying architectures, model-specific techniques that leverage internal structural knowledge, and inherently interpretable systems designed with transparency as a foundational principle. The article evaluates these methodologies against essential criteria, including fidelity, stability, user comprehensibility, and domain appropriateness, with special focus on highly regulated sectors where explanations are not optional but legally required. The article goes further into the new frontiers of investigation, including counterfactual expositions, causal interpretability frameworks, and the combination of explainability with fairness aspects. The field has made considerable advances, but it still struggles to standardize measurement of evaluation, deal with vulnerability to adversarial manipulation, and reconcile technical explanations with human cognitive patterns, which indicates a direction towards finding the middle ground between mathematical correctness and practical access to a wide range of stakeholders.

Keywords: Explainable Artificial Intelligence, Model Interpretability, Counterfactual Explanations, Causal Reasoning, Regulatory Compliance

1. Introduction

The resulting ironic paradox of the dramatic increase in artificial intelligence powers, and most notably deep learning systems, is that greater predictive accuracy is achieved at the expense of reduced interpretability. This black box phenomenon poses significant challenges in areas where transparency is a pillar of regulatory compliance, moral execution, and trust between stakeholders.

Explainable AI (XAI) was created specifically to work out of this dilemma, including diverse methodological directions that seek to make AI systems understandable without performance loss. According to Ribeiro and others, model decisions remain frequently incomprehensible to non-technical stakeholders, as machine learning experts continually struggle to explain their choices in practice [1]. The landscape has transformed markedly throughout recent years, with scholarly innovations spanning post-hoc explanation frameworks for existing opaque models to inherently transparent architectures engineered with clarity as a fundamental design principle.

Research from Lundberg and Lee revealed that model interpretability profoundly shapes user confidence, with experimental data indicating marked increases in trust levels when participants received coherent explanations for algorithmic determinations [2]. This finding highlights both technical and human-focused dimensions of the XAI challenge. Meanwhile, regulatory frameworks are becoming more and more explainable, as seen in the provisions of the European Union General Data Protection Regulation that provide specific explainability rights under the guise of a right to explanation, which impacts many thousands of organizations deploying AI tools across national borders.

With increasingly advanced AI architectures, which add parameters in the billions, the complexity of their decision-making processes grows exponentially more difficult to unravel. Evidence suggests that current post-hoc explanation techniques demonstrate limited fidelity when applied to intricate neural

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

architectures, pointing toward fundamental constraints within existing approaches. This problematic gap has catalyzed the proliferation of XAI research publications, reflecting widespread recognition that explainability constitutes an indispensable element of responsible AI deployment practices.

2. The Evolution of XAI Methods

2.1 Model-Agnostic Explanation Methods

Model-agnostic techniques provide versatility through explanations applicable to any machine learning model, regardless of internal architecture. These approaches conceptualize the underlying model as an impenetrable black box while focusing on input-output relationship analysis through various approximation strategies.

LIME (Local Interpretable Model-agnostic Explanations) approximates sophisticated models locally using simpler, inherently interpretable structures such as linear regression or decision trees [3]. Through systematic input perturbation around specific instances and monitoring corresponding outputs, LIME generates locally faithful explanations highlighting features with the greatest influence on particular predictions. Field evaluations demonstrate LIME's effectiveness across diverse applications, including textual classification, visual recognition, and structured data analysis.

SHAP (Shapley Additive exPlanations) consolidates multiple explanation frameworks within a unified theoretical structure based on game-theoretic Shapley values [4]. This methodology assigns importance values to individual features representing their contribution toward specific predictions, thereby providing both localized explanations for single instances and broader insights into overall model behavior. SHAP values provide mathematical assurances regarding local accuracy and consistency absent from alternative methods, rendering them particularly suitable for critical applications.

Anchors deliver explanations via decision rules that "anchor" predictions with exceptional precision. Unlike LIME's continuous approximation approach, Anchors presents discrete, rule-based explanations that any user can find intuitive and actionable. User evaluation studies indicate that rule-based explanations demonstrate superior comprehensibility among non-specialists compared to feature importance visualizations.

2.2 Model-Specific Techniques

Unlike model-agnostic methods, model-specific methods use the information that the model's internal structure is known, resulting in more accurate information, although at the cost of diminished predictive ability.

Saliency maps emphasize regions within input data (especially images) that most strongly influence model predictions. These visualizations help users grasp what neural networks "perceive" during decision processes, though recent literature questions their reliability. Studies identify troubling instability issues in conventional saliency methods, where minor input perturbations sometimes produce dramatically divergent explanations.

Grad-CAM (Gradient-weighted Class Activation Mapping) enhances saliency mapping by utilizing gradients flowing into final convolutional layers to produce coarse localization maps highlighting prediction-relevant regions. This technique proves particularly valuable for explaining convolutional neural networks in computer vision contexts. Medical imaging specialists extensively adopt Grad-CAM for validating diagnostic models by ensuring focus on clinically significant image regions.

Integrated Gradients addresses certain limitations of traditional gradient-based approaches by considering complete gradient paths from baselines (such as blank images) to inputs being explained. This methodology satisfies desirable theoretical properties, including sensitivity and implementation invariance. Organizations deploying large-scale vision models have incorporated integrated gradients into explanation frameworks based on these theoretical advantages.

2.3 Inherently Interpretable Architectures

Rather than retroactively explaining complex black-box models, certain researchers advocate designing inherently interpretable models from conception.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Decision trees and rule-based systems provide natural interpretability through hierarchical structures and explicit decision rules. While traditionally limited when handling complex data, recent advances in ensemble methodologies and optimization techniques have enhanced competitive performance. Financial sector organizations continue to favor these approaches for credit decisioning due to explicit logical structure and regulatory acceptance.

Generalised Additive Models (GAMs) are linear models that allow the use of non-linear association between single predictors and outcomes and additive structure. This design allows one to visualize the contribution of each feature separately, hence understanding complex relationships that would not be at the cost of predictive ability. Healthcare researchers successfully apply GAMs to patient risk stratification scenarios where interpretability proves crucial for clinical adoption.

Neural Additive Models (NAMs) and Explainable Boosting Machines (EBMs) represent newer developments combining modern machine learning predictive power with classical statistical model interpretability. Comparative analyses demonstrate these approaches achieve accuracy comparable to black-box neural networks while preserving interpretability for structured data tasks prevalent across regulated industries.

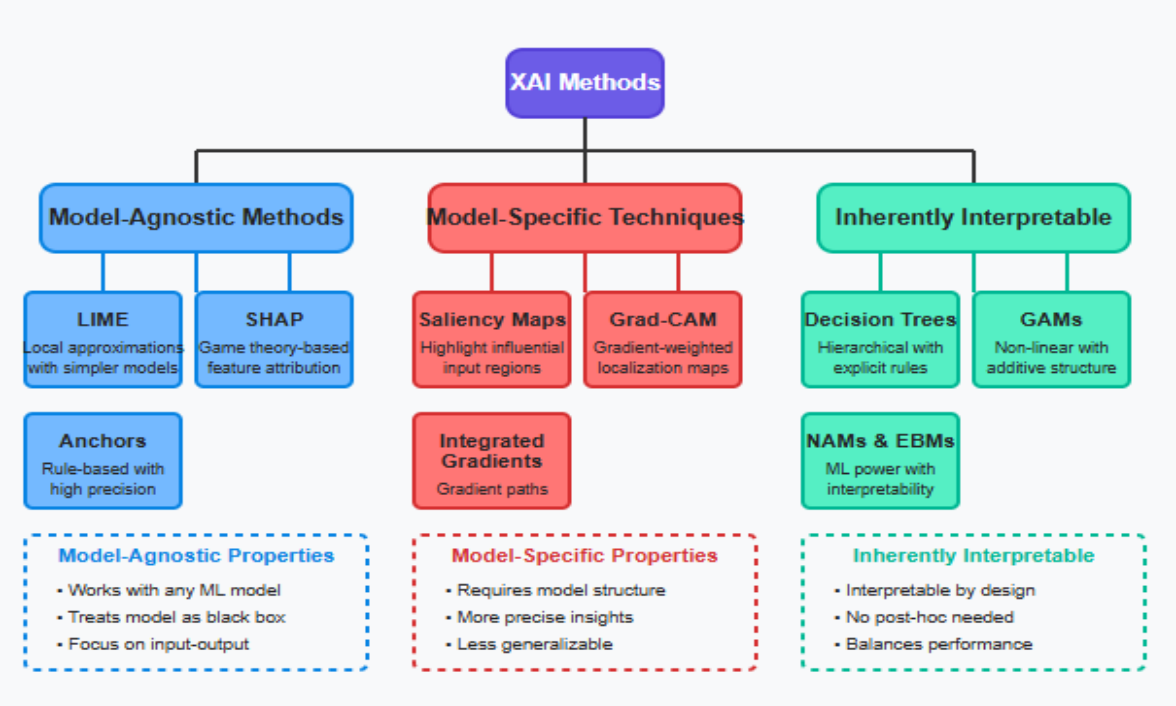

Fig 1: The Evolution of XAI Methods Classification [3, 4]

3. Evaluation Criteria for XAI Methods

Assessing explanation quality remains perhaps the most vexing challenge within XAI research. Four critical criteria have emerged from extensive fieldwork:

Fidelity gauges explanation accuracy relative to actual model behavior [5]. Truly high-fidelity explanations authentically reflect underlying model operations, whereas low-fidelity alternatives risk fundamentally misleading stakeholders about decision mechanisms. Extensive testing reveals that post-hoc explanation techniques frequently struggle when confronting sophisticated deep learning architectures. This creates a paradoxical tension: precisely those models demanding the greatest explanation due to complexity prove most resistant to reliable explanation through current techniques. Rigorous benchmark evaluations expose troubling discrepancies between explanation-indicated factors and genuine model behavior, particularly where models leverage subtle feature interactions that simplistic linear approximations cannot adequately capture.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Stability examines explanation consistency across similar inputs. Explanations exhibiting significant variation despite minimal input changes substantially erode trust and practical value. Deep investigations into gradient-based explanation methodologies have uncovered disturbing instability patterns, imperceptible input alterations sometimes radically transform resulting explanations without affecting predicted outcomes. Beyond mere usability complications, this vulnerability introduces security risks through potential adversarial manipulation, enabling malicious actors to conceal biased decision criteria while preserving problematic outcomes. Proposed countermeasures include specialized regularization techniques enhancing explanation stability alongside ensemble approaches aggregating multiple explanation methodologies.

User comprehensibility evaluates whether target audiences can meaningfully understand provided explanations [6]. It is critical to find a proper balance between technical sophistication and accessibility, but the explanations that show mathematical grace but are otherwise not comprehensible by the target users are technically flawed in their mission statements. Research into human explanation processing by cognitive scientists indicates that there are consistent preferences: explanations must be contrasting (to clarify why one of a range of things happened instead of the other), selective (emphasizing what is important about the situation and not everything), and socially aware (recognizing what is known in common by the explainer and the recipient). XAI systems that apply these principles show levels of user satisfaction and decision quality that are measurably better, which is why it is important to consider the alignment of the explanation design with the basic human cognitive patterns.

Domain suitability recognizes that effective explanation varies substantially across application contexts. Formats that are ideally adapted to medical diagnosis situations might not be suitable at all in financial fraud detection or content recommendation systems. Explanations that conform to proven clinical knowledge and causal relationships are usually highly valued by healthcare practitioners, whereas statistical reliability and compliance with regulations are highly valued by financial analysts. The differentiated patterns of preference are always identified in cross-domain user studies and require interpretation methods to be evaluated not only on technical measures but also on compatibility with domain needs. A growing consensus supports flexible XAI frameworks capable of adapting explanation approaches based on both technical model characteristics and contextual application requirements.

Criterion	Description	Key Challenges	Potential Solutions
Fidelity	Accuracy of explanation relative to actual model behavior	Post-hoc methods struggle with complex architectures; Linear approximations fail to capture subtle feature interactions	Hybrid explanation approaches; Model-specific techniques
Stability	Consistency of explanations across similar inputs	Gradient-based methods show vulnerability to minor input alterations; Security risks from adversarial manipulation	Specialized regularization techniques; Ensemble approaches aggregating multiple methods
User Comprehensibility	Whether target audiences understand explanations	Balancing technical sophistication with accessibility	Explanations that are contrastive, selective, and socially aware
Domain Suitability	Appropriateness for specific application contexts	Different domains require different explanation formats	Flexible XAI frameworks adapting to both technical model characteristics and contextual requirements

Table 1: Key Evaluation Dimensions for XAI Method Assessment [5, 6]

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

4. XAI in Regulated Domains

The issue of explainability is especially important in the context of high-stakes areas when algorithmic decision-making has a significant effect on human life and regulatory regulation directly requires such transparency.

4.1 Healthcare

Healthcare environments increasingly deploy AI systems supporting diagnostic processes, treatment recommendations, and risk stratification [7]. Resources like MIMIC-IV datasets provide invaluable foundations for developing and testing healthcare-focused explainable approaches. Beyond mere regulatory adherence, effective healthcare explanations must carefully balance technical precision with clinical relevance, presenting information complements rather than complicating medical decision processes.

Deep-learning-based clinical decision support systems are shown to possess impressive predictive power in a variety of medical fields, such as radiology or dermatology. But broad adoption has been hampered by ongoing resistance on grounds of the black box nature. Studies examining healthcare practitioner attitudes reveal consistent resistance toward implementing AI recommendations absent comprehensible reasoning explanations, regardless of documented performance metrics. Such opposition is an appropriate medical ethic and patient safety issue and not a mere professional pride.

The AI medical device regulatory structures that the FDA implements are more focused on the requirements of transparency and explainability. Recent guidance specifically mandates "human interpretable" model behavior representations, particularly for systems where incorrect decisions potentially cause serious patient harm. Successful healthcare XAI implementations typically feature multilayered explanation approaches, providing feature importance metrics for technical validators alongside conceptual, domain-appropriate explanations referencing familiar medical concepts and causal relationships for practicing clinicians.

4.2 Financial Services

Banking institutions that use AI to conduct credit score, fraud prediction, and investment suggestions have to meet tough regulatory transparency and fairness criteria. Approaches to explaining datasets such as those available in IEEE-CIS Fraud Detection benchmarks should address both technical performance concerns and adherence to regulatory requirements such as the Equal Credit Opportunity Act and GDPR right to explanation requirements.

Financial services present distinctive XAI implementation challenges through combined high regulatory scrutiny and sophisticated mathematical modeling. Credit scoring models must simultaneously predict default risk accurately while providing specific adverse decision justifications in consumer-accessible language. This dual requirement has driven innovative approaches bridging complex model internals and consumer-facing explanations.

Leading financial institutions report substantial XAI integration benefits, including reduced compliance costs, enhanced model governance, and improved customer satisfaction. Through enabling detailed model auditing capabilities, explainable approaches facilitate the identification and mitigation of potential bias before its manifestation in lending decisions. Industry-leading organizations have implemented comprehensive XAI frameworks spanning complete model lifecycles from development through deployment and ongoing monitoring.

4.3 Criminal Justice

Few AI applications raise more profound ethical questions than criminal justice implementations [8]. The controversial COMPAS recidivism prediction tool is the subject of far-reaching debates on fairness, bias, and transparency of algorithmic decision-making. The arguments in this context have to consider both technical performance issues and more general values in society with regard to justice, equity, and procedural fairness.

The analyses of the COMPAS system have proved some alarming differences in false positive rates of different demographic groups, which leads to the heated debate in the framework of algorithmic fairness in criminal justice settings. Subsequent investigations demonstrated that many issues

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

stemmed directly from insufficient transparency regarding factor weighting methodologies and historical data influences. Without adequate explanations, judges and stakeholders lacked the capacity to effectively evaluate whether system recommendations aligned with fundamental legal fairness and proportionality principles.

Multiple jurisdictions have responded by implementing strict explainability requirements for criminal justice algorithmic tools. These regulations typically mandate both technical transparency and explanation accessibility for defendants and legal representatives. Progressive approaches require explanation method validation through combined technical evaluation and stakeholder user studies, including judges, attorneys, and community representatives from affected populations.

COMPAS and similar system experiences highlight a crucial XAI insight within high-stakes domains: explanations must address not merely how systems function technically but whether that functioning aligns with domain-specific values and ethical principles. Technical accuracy alone proves insufficient when explanations cannot meaningfully integrate with existing decision processes and institutional frameworks.

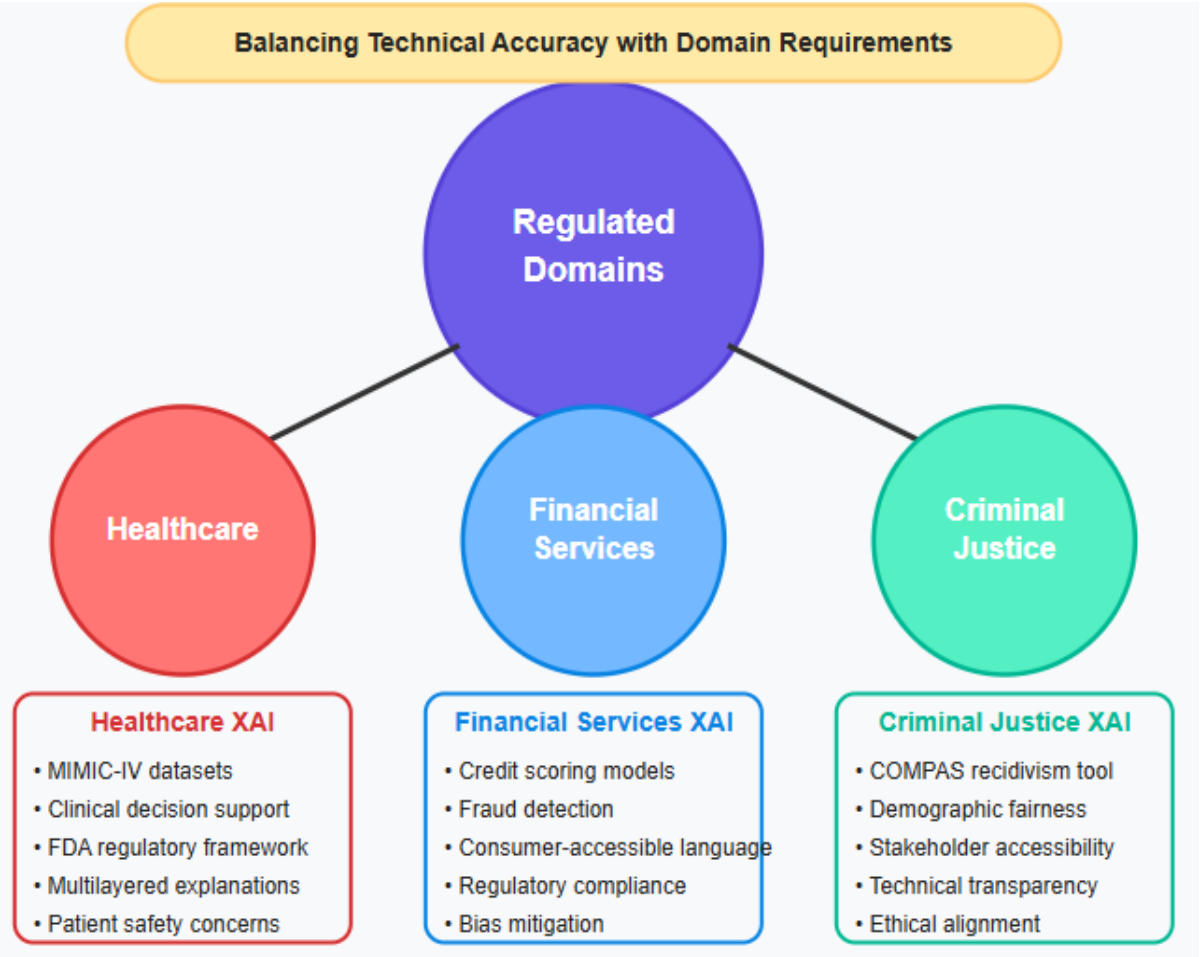

Fig 2: XAI Applications in Regulated Domains [7, 8]

5. Emerging Frontiers in XAI Research

Several promising research directions currently expand explainable AI boundaries, pushing beyond conventional approaches toward more sophisticated frameworks addressing fundamental limitations in existing methodologies.

Counterfactual explanations tackle the essential question "What changes would yield a different outcome?" [9]. These explanations deliver actionable insights by identifying minimal alterations

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

necessary for prediction changes, proving particularly valuable where stakeholders require guidance toward achieving desired outcomes.

Counterfactual approaches align remarkably well with cognitive science findings, indicating humans naturally reason through contrasting scenarios when understanding causality. While traditional feature importance methods merely highlight factors influencing current predictions, counterfactuals chart pathways toward alternative outcomes. This actionable characteristic proves especially beneficial within decision-support environments where understanding pathways to different results becomes critical.

Financial service providers have pioneered practical counterfactual explanation implementations, notably within credit application contexts. When rejecting credit applications, counterfactual systems generate individualized recommendations such as "Reducing debt-to-income ratio by just 4 percentage points would result in application approval." Such explanations simultaneously satisfy regulatory mandates while delivering practical guidance. Recent counterfactual generation algorithmic advances have substantially improved computational efficiency while strengthening feasibility constraints, ensuring recommended changes remain within realistic parameters.

Counterfactual explanation framework research has expanded toward addressing multiple simultaneous objectives, carefully balancing explanation simplicity, proximity to original instances, and alternative diversity. Healthcare applications exemplify this multi-objective approach, requiring counterfactuals that demonstrate not only accuracy but medical plausibility aligned with realistic treatment options. Surgical risk prediction models utilizing counterfactual explanations demonstrate measurably improved physician acceptance rates alongside enhanced patient comprehension compared with traditional explanation methodologies.

Causal interpretability transcends correlational explanations by addressing fundamental causal relationships within model decisions [10]. Through causal reasoning incorporation, these approaches strive toward explanations better aligned with human understanding regarding why events occur rather than merely identifying data patterns.

Causal interpretability frameworks represent nothing short of a paradigmatic shift within XAI conceptualization. While conventional explanation methods merely describe statistical relationships within model behavior, causal approaches attempt to uncover underlying mechanisms driving predictions. This distinction becomes absolutely critical within complex domains where spurious correlations potentially mislead models. Research combining structural causal modeling with deep learning architectures demonstrates how causal knowledge enhances both model performance and explanation quality.

Healthcare applications particularly highlight causal interpretability value, where treatment recommendations must consider intervention effects beyond mere statistical associations. Systems incorporating causal knowledge graphs demonstrate measurably improved capabilities, distinguishing genuine causation from mere correlation within clinical datasets, yielding substantially more reliable decision support. These approaches typically blend domain expertise with data-driven methodologies, creating explanations referencing established causal mechanisms familiar to practicing clinicians.

Counterfactual consistency, ensuring explanations maintain validity across different hypothetical scenarios, has emerged as a key research focus within causal interpretability. Methods addressing this challenge frequently incorporate formal causal inference techniques, including do-calculus for reasoning about intervention effects. Though computationally demanding, these approaches generate explanations maintaining logical consistency even when examining multiple potential interventions, representing a critical requirement within high-stakes decision support contexts.

Integration with fairness frameworks acknowledges fundamental interconnections between explanations and fairness considerations. Transparent systems facilitate bias detection, while fairness considerations profoundly shape what constitutes adequate explanation within contexts where equity remains paramount.

Research examining this intersection demonstrates explanation methods functioning as powerful tools for identifying and mitigating AI system bias. By exposing how models weigh various features,

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

particularly those correlating with protected attributes, explanations enable more effective fairness interventions. Conversely, fairness constraints guide explanation method development, highlighting problematic patterns rather than obscuring them.

Financial lending offers a compelling case study regarding this integration. Organizations implementing both XAI and algorithmic fairness frameworks report enhanced capabilities in identifying and addressing disparate impact before deployment. These integrated approaches typically combine technical solutions with governance processes, ensuring explanations revealing potential bias trigger appropriate review and mitigation procedures.

Emerging consensus strongly suggests that explanation and fairness should not represent separate concerns but rather complementary aspects within responsible AI development. This integrated perspective has substantially influenced regulatory frameworks and industry standards, with recent guidelines emphasizing explanation roles within fairness assessments and bias mitigation strategies.

Research Direction	Key Focus	Application Areas	Benefits	Challenges
Counterfactual Explanations	"What changes for a different outcome?"	Financial Services, Healthcare	Actionable guidance aligns with human reasoning	Balancing simplicity, proximity, and diversity
Causal Interpretability	Underlying causal mechanisms	Healthcare, Complex domains	Better alignment with human understanding, Improved reliability	Computational demands, Complex implementation
Fairness Integration	Interconnection between explanations and equity	Financial lending, Regulated domains	Bias detection and mitigation, Regulatory compliance	Balancing multiple objectives, Governance complexity

Table 2: Innovative Approaches in Modern Explainable AI Research [9, 10]

Conclusion

Explainable AI is the intersection of technical innovation and the principle of human-centered design, responding to the underlying necessity to make ever-more complex AI systems understandable by the people who utilize them and are impacted by them. As artificial intelligence proceeds in its transformation of industries and societies, the ability to offer meaningful explanations to the decisions made by algorithms rises not only as a technical question but also as an ethical requirement. The field has now advanced beyond crude post-hoc explanation techniques to advanced models that can take into account causal relationships, counterfactuality, and implications of fairness, but still cannot produce explanations that are both true to the complicated models and comprehensible to their different stakeholders. To proceed into the future, interdisciplinary cooperation in which machine learning researchers and domain experts, cognitive scientists, and regulatory experts work together to create a hybrid solution is essential that allowing balancing of technical values with practical use. In such joint endeavors, explainable AI can transform itself out of a technical niche to be an indispensable part of responsible AI development and implementation, trust-building, and fulfilling both compliance imperatives and moral duty in an ever-AI-influenced world.

Journal of Information Systems Engineering and Management
2025, 10(59s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

References

[1] Marco Tulio Ribeiro et al., "' Why Should I Trust You?': Explaining the Predictions of Any Classifier," KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. https://doi.org/10.1145/2939672.2939778
[2] Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," 31st Conference on Neural Information Processing Systems (NIPS 2017), 2017. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
[3] Christoph Molnar, "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable," 2019. https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf
[4] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic Attribution for Deep Networks," arXiv:1703.01365, 2017. https://arxiv.org/abs/1703.01365
[5] Mengjiao Yang and Been Kim, "Benchmarking Attribution Methods with Relative Feature Importance," arXiv:1907.09701, 2019. https://arxiv.org/abs/1907.09701
[6] Zachary C. Lipton, "The Mythos of Model Interpretability," arXiv:1606.03490, 2017. https://arxiv.org/abs/1606.03490
[7] Ahmad Chaddad et al., "Survey of Explainable AI Techniques in Healthcare," Sensors, 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC9862413/
[8] Partnership on AI, "Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System,". https://partnershiponai.org/wp-content/uploads/2021/08/Report-on-Algorithmic-Risk-Assessment-Tools.pdf
[9] Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," arXiv:1711.00399, 2018. https://arxiv.org/abs/1711.00399
[10] Bernhard Schölkopf, "Causality for Machine Learning," arXiv:1911.10500, 2019. https://arxiv.org/abs/1911.10500