Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Explainable and Fair Deep Learning Models for Automated Decision-Making in High-Stakes Financial and Healthcare Systems

Amjad Ali
Department of Computer Science University of Shangla, KP Pakistan
Email: amjad.cs@uop.edu.pk

ABSTRACT

The deployment of deep learning models in high-stakes domains such as healthcare diagnostics, credit scoring, insurance underwriting, and fraud detection has raised critical concerns regarding transparency, trustworthiness, and equity. This review examines the dual imperatives of explainable artificial intelligence (XAI) and algorithmic fairness to mitigate the "black-box" limitations of neural networks in automated decision-making systems. Key XAI techniques post-hoc methods (SHAP, LIME, Integrated Gradients), attention visualization, counterfactual explanations, and inherently interpretable architectures (concept bottleneck models, prototype-based networks) are evaluated for their ability to provide human-understandable rationales without compromising predictive performance. Fairness frameworks address group and individual fairness through pre-processing (re-weighting, adversarial debiasing), in-processing (constraint optimization), and post-processing (threshold adjustment) strategies, with particular attention to protected attributes (race, gender, age) and proxy variables. Application-specific insights highlight successes in medical imaging (explainable CNNs for chest X-ray interpretation), financial risk models (fair GNNs for fraud detection), and claims processing. Regulatory landscapes (EU AI Act, US NIST AI Risk Management Framework) and emerging standards (FAIR principles, ISO/IEC 42001) are discussed alongside persistent challenges: trade-offs between accuracy, explainability, and fairness; scalability of explanations for large models; and robustness against adversarial attacks on interpretability. The synthesis underscores the necessity of hybrid human-AI governance models to ensure accountable, equitable, and clinically/financially reliable automated decisions.

Keywords: Explainable Ai, Algorithmic Fairness, Deep Learning, High-Stakes Decision-Making, Healthcare Ai, Financial Ai, Shap, Lime, Adversarial Debiasing, Counterfactual Explanations, Regulatory Frameworks, Trustworthy AI

Introduction:

The contemporary landscape of artificial intelligence is defined by a fundamental shift from an era of raw predictive performance to an era of algorithmic accountability. Historically, the evaluation of machine learning models focused almost exclusively on statistical benchmarks such as precision, recall, and F1-scores (Scientific Journal of Artificial Intelligence and Blockchain Technologies, 2025). However, as deep

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

learning (DL) architectures have moved from controlled research environments into the operational core of high-stakes sectors like healthcare and finance, these accuracy-centric metrics have proven insufficient (ManageEngine Insight, 2025;In high-stakes environments defined as domains where automated decisions carry profound consequences for human life, financial solvency, and social equity the "black-box" nature of modern neural networks has become a significant liability (Elamin, 2024).

The central issue is no longer merely whether a model can predict an outcome correctly, but whether human stakeholders can understand and trust the rationale behind that prediction. For instance, a medical diagnostic tool achieving 98 percent accuracy is of limited utility if a clinician cannot validate its findings against medical literature or clinical intuition (Ferrario et al., 2020). Clinicians remain legally and ethically responsible for patient outcomes; therefore, their reluctance to adopt AI systems often stems from an inability to inspect the reasoning behind recommendations (Waithira et al., 2025). Similarly, in the financial sector, black-box models that outperform traditional methods in credit scoring may inadvertently perpetuate systemic discrimination, leading to regulatory backlash and loss of consumer trust (Annual Review of Financial Economics, 2023).

Explainable Artificial Intelligence (XAI) and algorithmic fairness have emerged as the primary frameworks for addressing these challenges. XAI seeks to transform inscrutable deep learning models into "accountable partners" by providing human-understandable justifications for their outputs (Wilson, 2025). Concurrently, fairness methodologies focus on detecting and mitigating biases that may arise from historical data or improper algorithmic design (Du et al., 2023). Together, these pillars form the foundation for trustworthy AI, ensuring that automated systems operate transparently, equitably, and in compliance with emerging legal mandates such as the European Union's Artificial Intelligence Act and the General Data Protection Regulation (GDPR) (European Union, 2024).

The Technical Dilemma: Opaque Architecture versus Human Interpretability

Deep learning’s unprecedented success in processing high-dimensional data such as medical imaging and complex financial time series is inextricably linked to its architectural complexity (Eddington, 2025). Models like Convolutional Neural Networks (CNNs), Transformers, and Deep Neural Networks (DNNs) utilize millions of parameters to capture non-linear relationships that traditional linear models cannot detect (Journal of Risk and Financial Management, 2025).

The trade-off between accuracy and interpretability is a pervasive theme in current research. Simple models, such as decision trees or linear regression, are "ante-hoc" or inherently interpretable; their internal logic is directly inspectable by humans (Science Publishing Group, 2025). Conversely, as researchers move toward more complex architectures like vast Transformer models for text analysis or deep CNNs for diagnostics, the potential for maximal predictive accuracy increases, but the decision-making process becomes more opaque (Omarov, 2025).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 1. Comparative Landscape of Model Architectures and Interpretability

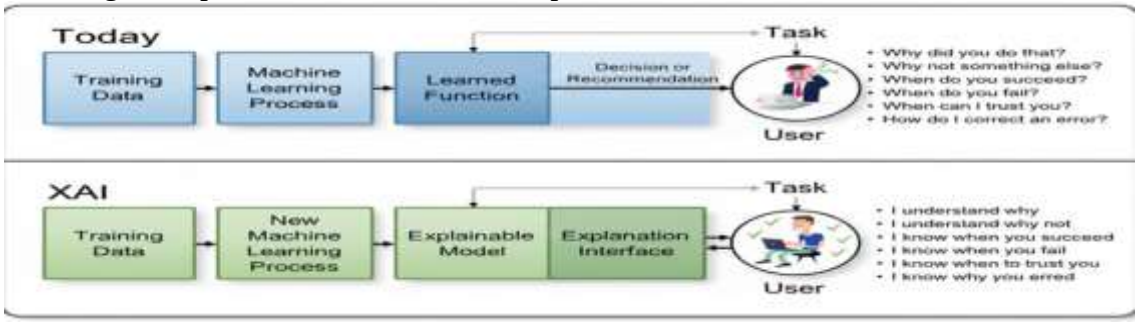
Architecture Class	Primary Mechanism	Predictive Power	Interpretability Level	Typical Domain
Linear Models	Weight-based linear sums	Moderate	Very High (Ante-hoc)	Credit scoring (Wilson, 2025)
Decision Trees	Hierarchical rule-based splits	Moderate	High (Ante-hoc)	Risk assessment (Wilson, 2025)
Ensemble Models	Voting across multiple learners	High	Low (Post-hoc needed)	Fraud detection (Agomuo et al., 2025)
CNNs	Spatial feature hierarchies	Very High	Opaque (Visualization needed)	Radiology (BMC Medical Imaging, 2026)
Transformers	Self-attention mechanisms	Extremely High	Opaque (Attention-based)	NLP, Financial series (Journal of Risk and Financial Management, 2025)
GNNs	Relational node embeddings	Very High	Opaque (Graph-explainer)	Insurance fraud (Zakaria et al., 2025)

Despite this apparent trade-off, some scholars critique the binary view that accuracy must be sacrificed for transparency. The "Rashomon effect" suggests that in many datasets, there may exist multiple models with nearly identical accuracy but varying levels of interpretability (Del Giudice, 2025). Reframing this dilemma as a strategic decision allows developers to prioritize inherently interpretable models in high-stakes settings where the risks of opacity outweigh the marginal gains in predictive performance (Rudin et al., 2022).

Taxonomy of Explainable AI (XAI) Methodologies

To bridge the cognitive gap between machine and human, the field of XAI has developed a diverse taxonomy of techniques, broadly categorized into ante-hoc and post-hoc methods (Cesarini et al., 2024).

Figure 1: Comparison of Traditional "Black-Box" Pipelines versus XAI Workflows featuring Transparent Models and User Explanation Interfaces.



Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Ante-hoc (Intrinsic) Interpretability

Ante-hoc models are "transparent by design". These models, such as RuleFit or sparse linear models, provide global explainability, offering a transparent view of how the system works across the entire feature space (Wilson, 2025). In regulated sectors like finance, ante-hoc models are often preferred for critical tasks like mortgage risk management where transparency is a non-negotiable regulatory requirement (Narteni, 2025).

Post-hoc (A posteriori) Explanations

Post-hoc methods are applied to black-box models after training to generate explanations for specific outputs. These are essential when the use of high-performance DL is unavoidable (Moradi & Samwald, 2021).

Feature Attribution Methods (SHAP and LIME): SHAP (SHapley Additive exPlanations) utilizes game theory to quantify the contribution of each feature to the final prediction (Devika et al., 2025). LIME (Local Interpretable Model-Agnostic Explanations) approximates the behavior of a complex model locally around a specific instance, making it useful for justifying individual loan denials or diagnostic predictions (Zafar & Khan, 2021).

Visual Explanations: Saliency maps and heatmaps highlight the regions of an input image or time-series data that most influenced the model's decision. Grad-CAM (Gradient-weighted Class Activation Mapping) is a dominant technique in radiology for visualizing the diagnostic rationale in CNNs (BMC Medical Imaging, 2026).

Counterfactual Explanations: These provide actionable insights by describing how a decision could have changed if certain inputs were different for example, informing a borrower that "If your income were 5,000 dollars higher, the loan would be approved" (Guidotti et al., 2019).

Attention Mechanisms: In Transformer models, attention maps visualize which parts of a sequence the model focused on, such as specific words in a clinical report or patterns in stock market data (Choi & Lee, 2023).

Dimensions of Algorithmic Fairness: Conceptual Frameworks and Metrics

In high-stakes decision-making, fairness is not a singular mathematical constant but a multidimensional construct that varies by application and jurisdiction (Hu et al., 2023). The literature distinguishes between different categories of harm that unfair AI can perpetuate: "allocative harm," where resources like credit are denied, and "quality of service harm," where the accuracy of services like medical diagnostics varies across demographics (Valentine et al., 2023).

Families of Fairness Metrics

To quantify discrimination, researchers utilize several mathematical definitions of fairness (Corbett-Davies et al., 2023).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Group Fairness (Parity-based): These metrics evaluate whether a model treats different demographic groups equally in its outcomes (Annual Review of Financial Economics, 2023).

Demographic Parity: Requires that the probability of receiving a positive outcome is independent of sensitive attributes. For example, the probability of approval for Group A equals the probability of approval for Group B, regardless of sensitive traits like race (Du et al., 2023).

Equalized Odds: A more refined metric ensuring that both the true positive and false positive rates are identical across groups (Machine Learning in Healthcare, 2023)

Individual Fairness: Grounded in the principle that similar individuals should receive similar outcomes. This requires a defined similarity metric between data points in the feature space (Srivastava et al., 2018).

Counterfactual Fairness: Assesses whether a decision would remain the same if an individual's protected attribute were changed while keeping other causal factors constant (Plecko and Bareinboim, 2024).

Table 2. Comparative Analysis of Fairness Metric Applications

Metric	Primary Sector	Implementation Goal	Notable Challenge
Demographic Parity	Public Policy / Hiring	Equalize representation across groups	May conflict with predictive accuracy (Du et al., 2025)
Equalized Odds	Healthcare Screening	Ensure equal diagnostic quality for all races	Requires ground truth labels for both groups (Machine Learning in Healthcare, 2023)
Equality of Opportunity	Financial Lending	Equalize true positive rates for qualified applicants	Disregards disparities in false positive rates (Du et al., 2025)
Predictive Parity	Risk Scoring (Recidivism)	Ensure equal precision across demographics	Hard to satisfy when base rates differ (Beyer Diaz et al., 2025; Journal of Risk and Financial Management, 2025)

Explainable and Fair Deep Learning in Healthcare Systems

The integration of AI into medicine offers transformative potential for diagnostics and resource allocation, yet it is fraught with risks related to "black-box" opacity and systemic bias (Panahi, 2025).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The Impact of Opaque Models on Clinical Adoption

Clinical decision-making requires transparency not only for trust but for legal and ethical defense. A study on sepsis prediction models highlighted that when clinicians lack visibility into how risk scores are calculated, they often ignore alerts, particularly when systems generate high rates of false alarms (Sendak et al., 2020). XAI transforms medical AI from an "inscrutable oracle" into a partner by providing saliency maps that allow doctors to verify whether an algorithm is focusing on relevant anatomical features or irrelevant image artifacts (John, 2025).

Figure 2: Multimodal XAI Framework for Healthcare: Connecting Clinicians, Patients, and Laboratory Services through Interpretable Decision Support.



Sources of Bias in Clinical Workflows

Algorithmic bias in healthcare arises from several systemic factors.

Data Acquisition Disparities: Institution-specific protocols can "leak" protected class information. For example, hospital-specific stain intensities in pathology datasets have been found to predict patient ethnicity, allowing models to use "shortcuts" for diagnosis rather than actual biological markers (Waithira et al., 2025).

Lack of Genetic Diversity: Many AI algorithms are trained on datasets over-representing individuals of European ancestry. This results in "population shifts" where performance degrades for East Asian or African populations, particularly in polygenic risk models (Özer et al., 2024).

Social Determinants of Health (SDoH): Inequities in access to care lead to "prevalence shifts," where certain demographics are diagnosed at later stages of a disease, causing AI models to associate those demographics with worse outcomes (Fabris et al., 2025).

Bias Mitigation in Medical Imaging

To address these issues, research into "disentanglement" and "representation learning" seeks to separate sensitive attributes from diagnostic features. Adversarial debiasing has also emerged as a powerful "in-processing" technique (Gori, 2024). By training a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

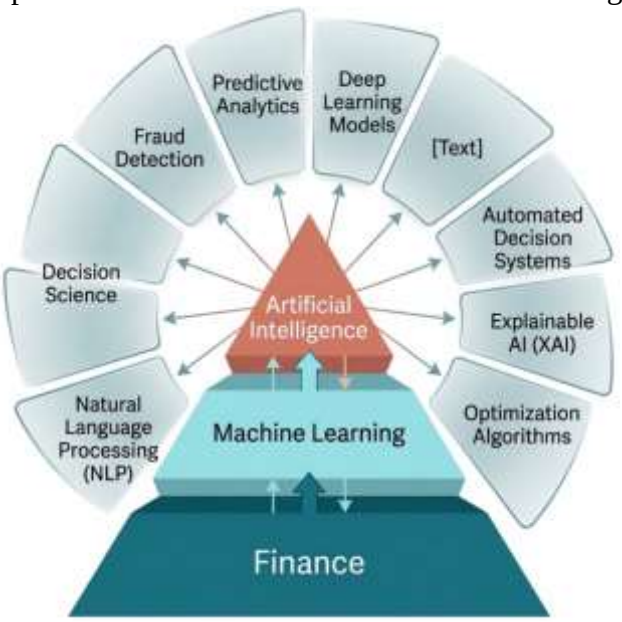
http://amresearchreview.com/index.php/Journal/about

predictor model alongside an adversarial network that attempts to detect sensitive attributes from the predictions, the main model is forced to learn features that do not correlate with protected status (Li et al., 2025). In diagnostic imaging, this has been shown to improve fairness metrics while maintaining clinically effective performance with negative predictive values greater than 0.98 (Machine Learning in Healthcare, 2023).

High-Stakes Decision-Making in Financial Systems

In the financial sector, deep learning is increasingly applied to credit scoring, fraud detection, and algorithmic trading. The complexity of financial data necessitates robust XAI to meet the strict demands of regulators and risk managers (Oko-Odion, 2025).

Figure 3: Conceptual Hierarchy of Artificial Intelligence, Machine Learning, and Specialized Sub-domains in Financial Technology.



Credit Scoring and the Compliance Trilemma

Lending institutions must balance three competing priorities: predictive accuracy, explainability, and fairness. While deep learning can achieve record-breaking precision in distinguishing between defaulters and non-defaulters, its opacity conflicts with the "Right to Explanation" under various consumer protection laws (Lending by Algorithm, 2025).

A common source of bias in financial AI is the use of "proxy variables". Even when protected attributes like race are removed from the dataset a method known as "Fairness through Unawareness" the model can still identify correlations with variables like ZIP codes or education levels to replicate discriminatory patterns (Chen et al., 2019).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Fraud Detection and Relational Data

The detection of medical insurance fraud or financial crime involves analyzing complex interactions between entities such as providers, patients, and banks. Graph Neural Networks (GNNs) are particularly effective here, modeling these interactions as a heterogeneous graph (Zakaria et al., 2025).

Heterogeneous Architectures: Models like HINormer and HybridGNN have shown superior performance in identifying abnormal billing patterns (Muhammad et al., 2025).

Graph Explanations: XAI techniques like GNNExplainer help analysts understand which specific nodes or edges (e.g., a cluster of suspicious transactions) led to a fraud flag. User studies indicate that XAI-augmented GNNs can increase analyst trust by 45 percent and reduce decision-making time by 35 percent compared to black-box models (Nandan et al., 2025).

Table 3. Performance Benchmarks for Hybrid and Graph-Based Models

Model Type	Application	Key Metric	Score/Performance	Feature Importance Tool
Hybrid (RF + XGB)	Real-time Fraud	Accuracy	97.1%	SHAP (Agomuo et al., 2025)
HINormer	Insurance Fraud	F-score	82.0% - 84.0%	GNNExplainer (Zakaria et al., 2025)
Random Forest	Credit Default	F1 Macro	90.0%	Permutation Importance (Lending by Algorithm, 2025)
SpikeNet (CNN)	MRI/Ultrasound	Accuracy	97.12% plus or minus 0.63%	XAlign (BMC Medical Imaging, 2026)
Log. Regression	Crime Detection	ROC AUC	0.856	Kernel Transform (Applied Machine Learning for Financial Fraud Detection, 2010)

Regulatory and Legal Landscapes of Automated Decision-Making

The global regulatory environment is rapidly evolving to mandate transparency and fairness in AI systems (Kashefi et al., 2024).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The GDPR and the Right to Explanation

The General Data Protection Regulation (GDPR) establishes the "Right to Explanation" for decisions based on automated processing that produce "legal or similarly significant effects". This necessitates that AI mechanisms be transparent and provide understandable decision-making methodologies. However, legal scholars debate whether this right applies to the "logic of the system" or the "specific outcome" for an individual, a distinction that fundamentally alters the technical requirements for XAI (Herrera, 2025).

The EU Artificial Intelligence Act

Entered into force in August 2024, the EU AI Act classifies AI according to risk levels

High-Risk Category: Includes AI for medical purposes and creditworthiness evaluation (except fraud detection) (Adenuga et al., 2023).

Article 86: Grants affected individuals a right to "clear and meaningful explanations" of the AI system's role in the decision-making process (Oko-Odion, 2025).

Provider Obligations: Developers of high-risk systems must establish risk management systems, ensure data governance to minimize bias, and provide documentation for human oversight (Dey et al., 2025).

Regional Variations: Law 5 and Bill C-27

Quebec's Law 5 is a pioneering example of sector-specific legislation, encoding the right to explanation in a healthcare context. This law mandates that patients be informed of the reasoning behind automated decisions, altering the physician's traditional "duty to inform" (European Union, 2024). However, the clause "based exclusively on automated processing" is a point of legal contention; if a human-in-the-loop (HITL) exists, the right to explanation might not be triggered, potentially exempting most current medical AI uses (Dubal, 2024).

Methodological Integration: Bridging Theory and Practice

Successfully deploying XAI and fair DL requires a multi-stage approach throughout the machine learning lifecycle (Radha et al., 2025).

Bias Mitigation Strategies

Pre-processing: Focuses on rebalancing raw data before training. This includes "Fair Data Transformers" and oversampling underrepresented groups using techniques like SMOTE (Sklavenitis et al., 2025).

In-processing: Integrates fairness directly into the model's objective function. Adversarial debiasing treats fairness as a game between a predictor and an adversary (Senevirathna et al., 2024).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Post-processing: Adjusts model outputs after training to correct residual biases. Methods like "multi-accuracy adjustments" refine treatment or loan recommendations without requiring the model to be retrained (Waithira et al., 2025).

The Role of User Interfaces in XAI

The value of an explanation is inextricably linked to its communicability. Effective XAI implementation requires "Explanation Interfaces" tailored to specific stakeholders (Wilson, 2025).

Interactive Dashboards: Visualizing key decision factors for technical teams (Del Giudice, 2025)

Plain-Language Summaries: Explaining credit or medical decisions in terms understandable to the general public (Eddington, 2025).

Expert Validation: Trust scoring where domain experts (e.g., radiologists) validate that the model's "attention" aligns with clinical logic (Scientific Journal of Artificial Intelligence and Blockchain Technologies, 2025).

Evaluation and Benchmarking: The Search for Reliable Standards

A critical gap in the field is the lack of standardized metrics for evaluating "explainability". While accuracy has objective scores, interpretability is subjective and context-dependent (Science Publishing Group, 2025).

Limitations of Current Benchmark Datasets

Many researchers rely on public benchmarks that have significant limitations for high-stakes research (Rudin et al., 2022).

German Credit and UCI Credit: These datasets are often too small and outdated for modern deep learning, containing fewer than 1,000 samples and lacking the complexity of contemporary financial markets (Özer et al., 2024).

MIMIC-III: While transformative for intensive care research, it lacks "longitudinal context" the multi-visit patient trajectories required to evaluate chronic disease management or care pathway optimization (Beyer Diaz et al., 2025).

Data Contamination: Public benchmarks often "leak" into the training corpora of large models (like GPT-4), leading to inflated performance scores and a false sense of model generalization (Cracks in the Current Paradigm, 2026).

Table 4. Evaluative Framework for Fairness and XAI Metrics

Metric Category	Specific Tool	Primary Evaluation Goal	Limitation
Faithfulness	Insertion/Deletion AUC	Measure if identified features	Computationally expensive for large models (BMC

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

		truly drive the decision	Medical Imaging, 2026)
Localization	Intersection over Union (IoU)	Check alignment with medical annotations (e.g., tumor masks)	Requires manual ground truth from experts (BMC Medical Imaging, 2026)
Stability	Explanation Variance	Ensure explanations don't change with small noise (Wilson, 2025)	Sensitive to perturbation parameters (Annual Review of Financial Economics, 2023)
Efficiency	Per-image Latency (ms)	Assess feasibility for real-time diagnostics	Often trades off with faithfulness (BMC Medical Imaging, 2026)
Human Trust	Likert Scale Surveys	Measure perceived transparency and reliability	Subject to cognitive biases (Scientific Journal of Artificial Intelligence and Blockchain Technologies, 2025; Zakaria et al., 2025)

The Socio-technical Trilemma: Accuracy, Fairness, and Explainability

The deployment of deep learning in high-stakes domains is not merely a technical challenge but a socio-technical one, where technical components and social structures jointly shape outcomes. Practitioners are often faced with a "Trilemma" where optimizing for one dimension may compromise others (Lending by Algorithm, 2025).

Path-Specific Excess Loss (PSEL)

Introducing fairness constraints often leads to an "excess loss" in predictive utility. Causal analysis indicates that enforcing group fairness (e.g., equalized odds) almost always results in a trade-off with accuracy, as the model is forced to ignore high-signal features that are causally linked to protected attributes (Plecko and Bareinboim, 2024). However, some multi-objective training approaches have demonstrated that by explicitly targeting both dimensions, models can be identified that satisfy both accuracy and fairness requirements more effectively than single-objective models (Fair-Eye Net, 2026).

The Fidelity Gap in Post-hoc XAI

While techniques like SHAP and LIME offer a way to interpret complex models, they are fundamentally "approximations" (Corbett-Davies et al., 2023). If the explanation model does not have high "fidelity" to the original black-box model, it may provide a false sense of security, misleading users into trusting a biased or erroneous system (Science Publishing Group, 2025). This has led many experts to argue that for the highest stakes such as terminal illness diagnosis or criminal sentencing black-box models should be banned in favor of inherently interpretable ones (Chen et al., 2019).

Online ISSN 3007-3197
Print ISSN 3007-3189

Conclusions

The transition of deep learning from laboratory benchmarks to operational deployment in healthcare and finance has exposed fundamental tensions between predictive power and societal accountability. Explainable AI and fairness methodologies are no longer optional enhancements but essential requirements for ethical, legally defensible, and clinically/financially trustworthy systems. Post-hoc and ante-hoc interpretability techniques have matured sufficiently to provide actionable insights for domain experts, while fairness interventions when applied thoughtfully across the pipeline can substantially reduce disparate impact without catastrophic performance degradation. Yet significant hurdles persist: the computational cost of high-fidelity explanations, the risk of superficial interpretability masking underlying biases, and the difficulty of defining context-dependent notions of fairness in heterogeneous populations. Regulatory momentum (EU AI Act, NIST AI RMF) and standardization efforts signal growing institutional recognition of these issues. The path forward lies in interdisciplinary collaboration integrating technical advances with ethical governance, continuous auditing, and participatory design involving clinicians, financial regulators, and affected communities to realize the full societal benefit of AI while minimizing harm. Only through such holistic accountability can automated decision-making systems earn legitimate trust in domains where errors carry profound human and economic consequences.

References

Devika, P., Kumar, G. M., Kumar, P. N., & Prabhu, S. N. (2025). XAI-powered hybrid model for real-time financial fraud detection. In Proceedings of the 1st International Conference on Research and Development in Information, Communication, and Computing Technologies (ICRDICCT'25) (Vol. 2, pp. 214–218). SCITEPRESS. https://doi.org/10.5220/0013880300004919

Annual Review of Financial Economics. (2023, November). Algorithmic fairness. Annual Review of Financial Economics, 15. https://doi.org/10.1146/annurev-financial-110921-125930

Applied Machine Learning for Financial Fraud Detection. (2010, August). Transactions of a customer over a time period are processed to form multivariate time series (Master's thesis).

Beyer Diaz, S., Coussement, K., & De Caigny, A. (2025). From collaborative filtering to deep learning: Advancing recommender systems with longitudinal data in the financial services industry. European Journal of Operational Research, 323(2), 609–625. https://doi.org/10.1016/j.ejor.2024.08.012

BMC Medical Imaging. (2026, January 5). Explainable Artificial Intelligence (XAI) is crucial for enhancing transparency and trustworthiness in medical imaging. BMC Medical Imaging.

Cracks in the Current Paradigm. (2026). Data contamination and risk of strategic cherry-picking. arXiv. https://arxiv.org/html/2510.07575v2

Deep Science Research. (2025). Hybrid models that combine symbolic AI with neural networks. Deep Science Research, 6(51).

Du, M., Cheng, L., & Dou, D. (2023). Algorithmic fairness. In L. Rokach, O. Maimon, & E. Shmueli (Eds.), Machine learning for data science handbook:

Online ISSN 3007-3197
Print ISSN 3007-3189

Data mining and knowledge discovery handbook (pp. 867–886). Springer. https://doi.org/10.1007/978-3-031-24628-9_45

European Union. (2024, August 1). EU Artificial Intelligence Act. https://artificialintelligenceact.eu/high-level-summary/

Fair-Eye Net. (2026). Fair-Eye Net: Uncertainty-aware hierarchical gating for intersectional fairness in clinical diagnostics. arXiv. https://arxiv.org/html/2601.18464v1

Journal of Risk and Financial Management. (2025). Deep learning and its applications in the financial industry. Journal of Risk and Financial Management, 19(2), 104. https://doi.org/10.3390/jrfm19020104

Lending by Algorithm. (2025, June). Lending by algorithm: Fair or flawed? An information-theoretic view of credit decision pipelines. Open Access Library Journal.

Machine Learning in Healthcare. (2023). Adversarial training improves outcome fairness. Scientific Reports.

ManageEngine Insight. (2025, September 12). Explainable AI in high-stakes decision environments. ManageEngine Insight. https://insights.manageengine.com/artificial-intelligence/explainable-ai-in-high-stakes-decision-environments/

Plecko, D., & Bareinboim, E. (2024). Path-specific excess loss (PSEL): A causal lens for the tension between fairness and accuracy. Scientific Reports.

Science Publishing Group. (2025). The enduring accuracy-interpretability trade-off in machine learning. Science Publishing Group. https://www.sciencepublishinggroup.com/article/10.11648/j.ajist.20250903.15

Scientific Journal of Artificial Intelligence and Blockchain Technologies. (2025, July). Explainable AI in high-stakes decision making: Beyond accuracy. Scientific Journal of Artificial Intelligence and Blockchain Technologies, 2(3).

Waithira, J., Chweya, R., & Cyprian, R. M. (2025). Adversarial debiasing for bias mitigation in healthcare AI systems: A literature review. Open Access Library Journal, 12(05), 1–13. https://doi.org/10.4236/oalib.1113340

Wilson, A. (2025, August 7). Explainable AI in finance. CFA Institute. https://rpc.cfainstitute.org/research/reports/2025/explainable-ai-in-finance

Zakaria, R. M., Rahman, M. M., Choudhury, M. T. H., Rahman, H., & Rafi, M. A. (2025). Detecting financial fraud in real-time transactions using graph neural networks and anomaly detection techniques. Journal of Economics, Finance and Accounting Studies, 7(6), 1–11. https://doi.org/10.32996/jefas.2025.7.6.1

Delipetrev, B., Tsinaraki, C., & Kostic, U. (2020). Historical evolution of artificial intelligence.

Elamin, M. O. I. (2024). AI through the ages: Unlocking key opportunities and navigating challenges in the history and future of artificial intelligence. International Journal of Religion, 5(12), 1152-1166.

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions. Philosophy & Technology, 33(3), 523-539.

Eddington, D. P. (2025). Unveiling Insights From Complexity: Advanced Computational Techniques for High-Dimensional Medical Data.

Online ISSN 3007-3197
Print ISSN 3007-3189

Omarov, B. (2025). Deep Learning in Biomedical Image and Signal Processing: A Survey. Computers, Materials, & Continua, 85(2), 2195.

Del Giudice, M. (2024). The prediction-explanation fallacy: A pervasive problem in scientific applications of machine learning. Methodology, 20(1), 22-46.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, 1-85.

Cesarini, M., Malandri, L., Pallucchini, F., Seveso, A., & Xing, F. (2024). Explainable ai for text classification: Lessons from a comprehensive evaluation of post hoc methods. Cognitive Computation, 16(6), 3077-3095.

Narteni, S. (2025). Reliability Assessment Methods for Explainable Artificial Intelligence.

Moradi, M., & Samwald, M. (2021). Post-hoc explanation of black-box classifiers using confident itemsets. Expert Systems with Applications, 165, 113941.

Zafar, M. R., & Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. Machine Learning and Knowledge Extraction, 3(3), 525-541.

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., & Turini, F. (2019). Factual and counterfactual explanations for black box decision making. IEEE Intelligent Systems, 34(6), 14-23.

Choi, S. R., & Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. Biology, 12(7), 1033.

Hu, W. C., Dillon, H. C., & Wilkinson, T. J. (2023). Educators as judges: Applying judicial decision-making principles to high-stakes education assessment decisions. Teaching and learning in medicine, 35(2), 168-179.

Valentine, N., Durning, S. J., Shanahan, E. M., & Schuwirth, L. (2023). Fairness in assessment: identifying a complex adaptive system. Perspectives on Medical Education, 12(1), 315.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., & Goel, S. (2023). The measure and mismeasure of fairness. Journal of Machine Learning Research, 24(312), 1-117.

Srivastava, M., Heidari, H., & Krause, A. (2019, July). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2459-2468).

Panahi, O. (2025). The Algorithmic Clinician: AI's Transformative Role in Modern Medicine. Int J Nurs Health Care, 2(1), 1-4.

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., ... & O'Brien, C. (2020, January). " The human body is a black box" supporting clinical decision-making with deep learning. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 99-109).

John, I. (2025). Regulatory and Ethical Implications of Advanced Predictive Models in Healthcare Decision-Making.

Özer, M., Perc, M., & Suna, H. E. (2024). Artificial intelligence bias and the amplification of inequalities in the labor market. Journal of Economy Culture and Society, (69), 159-168.

Online ISSN 3007-3197
Print ISSN 3007-3189

Fabris, A., Baranowska, N., Dennis, M. J., Graus, D., Hacker, P., Saldivar, J., ... & Biega, A. J. (2025). Fairness and bias in algorithmic hiring: A multidisciplinary survey. ACM Transactions on Intelligent Systems and Technology, 16(1), 1-54.

Gori, P. (2024). Modeling, Learning, and Transferring Anatomical Representations in Medical Imaging using AI (Doctoral dissertation, Institut polytechnique de Paris).

Li, J., Khayatkhoei, M., Zhu, J., Xie, H., Hussein, M. E., & AbdAlmageed, W. (2025). Fairness and Disentanglement: A Critical Review of Predominant Bias in Neural Networks. Transactions on Machine Learning Research.

Oko-Odion, C. (2025). Ai-driven risk assessment models for financial markets: Enhancing predictive accuracy and fraud detection. International Journal of Computer Applications Technology and Research, 14(04), 80-96.

Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019, January). Fairness under unawareness: Assessing disparity when protected class is unobserved. In Proceedings of the conference on fairness, accountability, and transparency (pp. 339-348).

Muhammad, R., Tbaishat, D., Nazir, A., Yacoub, S., AbdulRazek, M., El-Enen, M. A. A., & Sahlol, A. T. (2025). Fraud detection and explanation in medical claims using GNN architectures. Scientific Reports.

Nandan, M., Mitra, S., & De, D. (2025). GraphXAI: a survey of graph neural networks (GNNs) for explainable AI (XAI). Neural Computing and Applications, 37(17), 10949-11000.

Kashefi, P., Kashefi, Y., & Ghafouri Mirsaraei, A. (2024). Shaping the future of AI: balancing innovation and ethics in global regulation. Uniform Law Review, 29(3), 524-548.

Herrera, F. (2025). Making sense of the unsensible: Reflection, survey, and challenges for xai in large language models toward human-centered ai. arXiv preprint arXiv:2505.20305.

Adenuga, A. A., Gaffar, O., Sikiru, A. O., & Otunba, M. (2023). AI in Financial Services and Fraud Prevention: Enhancing Fraud Detection, Risk Management, and Creditworthiness Assessment. Gyanshauryam, International Scientific Refereed Research Journal, 6(2).

Oko-Odion, C. (2025). Ai-driven risk assessment models for financial markets: Enhancing predictive accuracy and fraud detection. International Journal of Computer Applications Technology and Research, 14(04), 80-96.

Dey, R., Roy, A., Akter, J., Mishra, A., & Sarkar, M. (2025). AI-driven machine learning for fraud detection and risk management in US healthcare billing and insurance. Journal of Computer Science and Technology Studies, 7(1), 188-198.

Dubal, V. (2024). Data Laws at Work. Yale LJF, 134, 405.

Radha, S. K., Kuehlkamp, A., & Nabrzyski, J. (2025). Advancing transparency and responsibility in machine learning: The critical role of FAIR principles-A comprehensive review. ACM Journal on Responsible Computing, 2(4), 1-38.

Sklavenitis, D., & Kalles, D. (2025). A Scoping Review and Assessment Framework for Technical Debt in the Development and Operation of AI/ML Competition

Online ISSN 3007-3197
Print ISSN 3007-3189

Platforms. Applied Sciences, 15(13), 7165.

Senevirathna, T., La, V. H., Marcha, S., Siniarski, B., Liyanage, M., & Wang, S. (2024). A survey on XAI for 5G and beyond security: Technical aspects, challenges and research directions. IEEE Communications Surveys & Tutorials, 27(2), 941-973.