Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Artificial Intelligence and Advanced Algorithms

*Review*

# Trustworthy Artificial Intelligence in Financial Decision-Making: A Systematic Review of Explainability, Fairness, and Accountability

**Minhao Li [1,*] and Shuyang Xu [2]**

[1] Master of Science in Computer Engineering, University of California, Davis, Davis, USA
[2] Master of Professional Studies, Applied Statistics, Cornell University, Ithaca, USA
* Correspondence: Minhao Li, Master of Science in Computer Engineering, University of California, Davis, Davis, USA

**Abstract:** The rapid adoption of artificial intelligence (AI) in financial services has introduced critical concerns regarding the transparency, equity, and governance of algorithmic decision-making. This paper presents a systematic review of 43 peer-reviewed studies published between 2018 and 2024, examining three core dimensions of trustworthy AI in finance: explainability, fairness, and accountability. The review synthesizes findings across credit scoring, fraud detection, risk management, and algorithmic trading domains. Results indicate that post-hoc explainability methods such as SHAP and LIME dominate current implementations, while fairness-aware approaches remain underexplored relative to performance optimization. A persistent trade-off between predictive accuracy and fairness is documented across multiple application contexts. This paper contributes a structured analytical framework and identifies gaps that warrant future investigation under evolving regulatory mandates including the EU AI Act.

**Keywords:** Trustworthy AI; Financial Decision-Making; Explainable AI; Algorithmic Fairness

## 1. Introduction

### *1.1. Research Background and Motivation*

The financial services industry has undergone a fundamental transformation driven by the integration of artificial intelligence into core operational processes. Credit scoring, fraud detection, portfolio optimization, and risk assessment now rely extensively on machine learning algorithms capable of processing high-dimensional data at scale. Lundberg and Lee demonstrated that Shapley value-based attribution methods can decompose complex model outputs into interpretable feature contributions, establishing a foundational tool adopted across financial applications [1].

Financial institutions worldwide deploy AI-powered scoring and detection mechanisms that affect the economic lives of millions of individuals. Bajracharya et al. surveyed recent developments in algorithmic biases within financial services, identifying that discriminatory patterns embedded in training data frequently propagate through model outputs, disproportionately affecting minority populations in lending and credit allocation contexts [2].

### *1.2. Research Questions and Objectives*

This review is guided by three research questions: (RQ1) What explainability methods are currently applied in financial AI and how effective are they? (RQ2) To what extent have fairness and bias mitigation strategies been integrated into financial AI

Received: 20 March 2026
Revised: 28 April 2026
Accepted: 09 May 2026
Published: 13 May 2026

applications? (RQ3) What accountability and governance frameworks exist to ensure responsible deployment of AI in financial services? The primary objective is to provide a structured synthesis of current knowledge across these three trustworthiness dimensions and to identify critical gaps in the literature.

### *1.3. Scope, Methodology Overview, and Contributions*

The scope of this review encompasses peer-reviewed conference and journal publications from 2018 to 2024, retrieved from IEEE Xplore, ACM Digital Library, Scopus, and Web of Science. Barocas et al. articulated a comprehensive theoretical framework connecting machine learning fairness to broader societal implications, and this review builds upon that foundation by situating explainability and accountability alongside fairness within financial decision-making [3]. The contributions are threefold: (a) a structured taxonomy of trustworthiness dimensions as they apply to financial AI, (b) a comparative analysis of explainability and fairness methods across domains, and (c) identification of regulatory alignment gaps.

## 2. Theoretical Foundations

### *2.1. AI Applications in Financial Decision-Making: An Overview*

AI-driven models have penetrated virtually every segment of the financial services value chain. Credit scoring constitutes one of the most extensively studied domains, where supervised learning algorithms predict borrower default probabilities based on historical repayment records and alternative data sources. Kozodoi et al. examined fairness in credit scoring through profit-adjusted assessment, demonstrating that incorporating fairness constraints need not result in prohibitive profitability reductions, though quantifiable trade-offs exist [4]. Fraud detection represents another domain where AI adoption has accelerated, with machine learning classifiers achieving recall rates exceeding 95% in controlled experimental settings.

### *2.2. Trustworthy AI: Definitions, Dimensions, and Taxonomies*

The concept of trustworthy AI encompasses multiple interdependent dimensions. Bracke et al. applied machine learning explainability to default risk analysis and identified that the capacity of a model to provide meaningful explanations is a necessary condition for institutional trust in automated lending decisions [5]. Arrieta et al. presented a comprehensive taxonomy of explainable AI covering concepts, opportunities, and challenges, classifying explanation methods along two primary axes: model-agnostic versus model-specific techniques, and global versus local explanation scopes [6].

Fairness in algorithmic decision-making has been formalized through multiple competing definitions. Verma and Rubin catalogued over twenty distinct mathematical formulations of fairness, including demographic parity, equalized odds, predictive parity, and calibration [7]. The mutual incompatibility of certain fairness definitions --- a result established by the impossibility theorem --- imposes structural constraints on any algorithmic system designed to satisfy multiple fairness criteria simultaneously. Li et al. advanced a holistic framework linking trustworthy AI principles to implementation practices, arguing that explainability, fairness, robustness, privacy, and accountability must be treated as interconnected rather than isolated requirements throughout the AI lifecycle [8].

### *2.3. Regulatory Landscape and Compliance Requirements for AI in Finance*

The regulatory environment governing AI in financial services has intensified since 2020. The EU AI Act, effective August 1, 2024, classifies credit scoring and fraud detection as high-risk AI applications subject to mandatory conformity assessments and human oversight. Non-compliance penalties can reach 7% of global annual turnover. In the United States, the CFPB has reinforced the requirement that lenders provide specific adverse action reasons for credit denials regardless of whether an AI model generated the decision.

## 3. Research Methodology

### *3.1. Systematic Literature Review Protocol Design*

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines adapted for interdisciplinary research. The protocol comprises five sequential phases: (1) formulation of research questions, (2) identification of relevant databases and search terms, (3) application of inclusion and exclusion criteria, (4) data extraction and quality assessment, and (5) thematic synthesis.

The initial search was conducted in December 2024 and yielded 1,247 candidate records. Ribeiro et al. introduced LIME as a model-agnostic local explanation method, and the conceptual foundations established in that work informed the construction of our search vocabulary spanning both the technical explainability literature and the financial services domain [9].

### *3.2. Search Strategy, Database Selection, and Inclusion/Exclusion Criteria*

Four electronic databases served as primary sources: IEEE Xplore, ACM Digital Library, Scopus, and Web of Science. The search query combined terms from three categories using Boolean operators. Category A included AI-related terms (artificial intelligence, machine learning, deep learning). Category B included financial domain terms (credit scoring, fraud detection, risk management, lending, algorithmic trading). Category C included trustworthiness terms (explainability, fairness, bias, accountability, transparency). The final query required at least one term from each category in the title, abstract, or keyword fields.

Inclusion criteria specified that studies must: (a) address at least one trustworthiness dimension in the context of financial AI, (b) be published between 2018 and 2024, and (c) be written in English. Exclusion criteria removed duplicates, non-empirical contributions, papers focused exclusively on performance without trustworthiness, and studies outside the financial domain.

Deck et al. conducted a critical survey on fairness benefits of explainable AI and noted that many claims regarding the XAI-fairness intersection lack empirical substantiation, reinforcing our decision to include only studies with verifiable contributions [10]. After deduplication, 843 unique records remained. Screening reduced this to 196 candidates, and full-text assessment produced a final set of 43 studies.

**Table 1.** Distribution of Reviewed Studies by Domain and Trustworthiness Dimension

| Application Domain | Explainability | Fairness | Accountability | Total | Proportion (%) |
|---|---|---|---|---|---|
| Credit Scoring | 13 | 8 | 4 | 17 | 39.5 |
| Fraud Detection | 10 | 3 | 2 | 12 | 27.9 |
| Risk Management | 5 | 4 | 3 | 8 | 18.6 |
| Algorithmic Trading | 2 | 2 | 1 | 4 | 9.3 |
| Insurance Underwriting | 1 | 1 | 0 | 2 | 4.7 |
| Total | 31 | 18 | 10 | 43 | 100.0 |

### *3.3. Data Extraction, Coding Scheme, and Quality Assessment*

A structured data extraction form captured the following attributes: publication year, venue type, financial application domain, trustworthiness dimension(s), methods employed, dataset characteristics, evaluation metrics, and key findings. Two independent reviewers coded each study, with inter-rater agreement measured using Cohen's kappa ($\kappa = 0.83$, indicating substantial agreement).

Molnar provided a comprehensive guide to interpretable machine learning, and the taxonomic categories proposed therein served as the basis for our coding scheme classifying explainability approaches into four groups: feature attribution methods (SHAP, LIME, permutation importance), rule-based methods (decision trees, rule extraction), example-based methods (counterfactual explanations, prototypes), and visualization-based methods (partial dependence plots, accumulated local effects) [11].

Quality assessment applied a modified Mixed Methods Appraisal Tool (MMAT). Each study was evaluated on five criteria: clarity of objectives, appropriateness of methodology, adequacy of experimental design, validity of analysis, and relevance of contributions. Studies scoring below 60% (3 of 46 initially screened) were excluded.

Figure 1 presents a stacked bar chart showing the temporal distribution of the 43 reviewed studies from 2018 to 2024. Each bar is divided according to the primary research focus, with explainability represented in blue, fairness (orange), and accountability (green). A secondary overlay line traces cumulative percentage growth. The visualization reveals a pronounced upward trajectory beginning in 2021, with the most substantial increase between 2022 and 2023 (a 47% rise). Explainability-focused studies dominate across all years, while fairness-oriented research shows notable growth only from 2022 onward. Over 65% of all reviewed studies were published in 2022--2024 (As shown in Table 2).

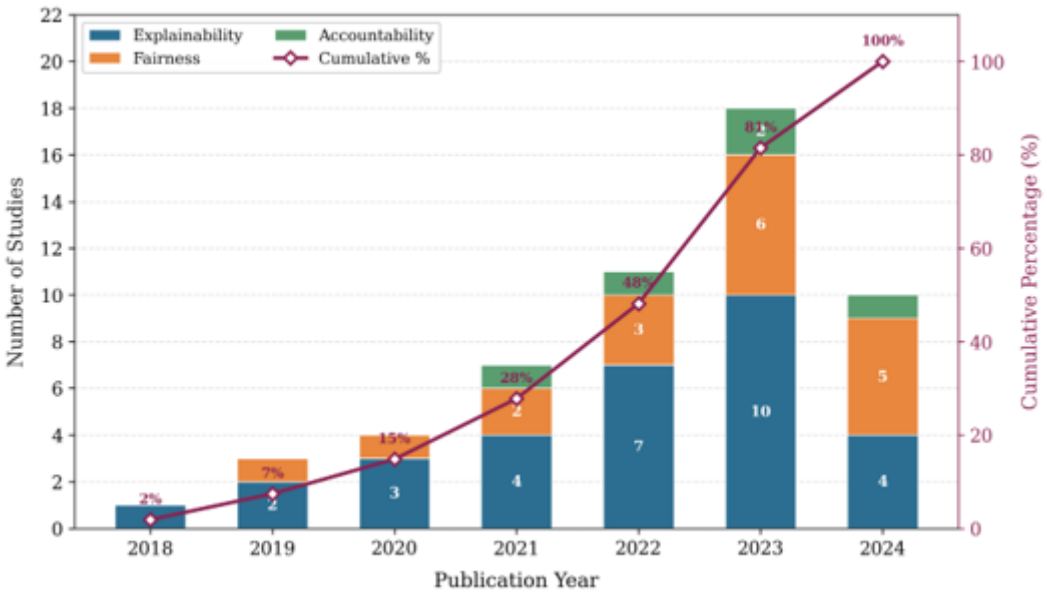


**Figure 1.** Distribution of Reviewed Studies by Publication Year and Research Focus

**Table 2.** Comparison of Post-Hoc Explainability Methods in Financial AI

| Method | Explanation Scope | Adoption Rate (%) | Avg. Fidelity | Cost | Primary Application |
|---|---|---|---|---|---|
| SHAP | Global + Local | 67.7 | 0.91 | High | Credit Scoring |
| LIME | Local | 45.2 | 0.84 | Medium | Fraud Detection |

| | | | | | |
|---|---|---|---|---|---|
| Permutation Imp. | Global | 29.0 | 0.78 | Low | Risk Management |
| PDP | Global | 22.6 | 0.73 | Low | Credit Scoring |

## 4. Results and Analysis

### *4.1. Explainability Methods in Financial AI: Current Approaches and Effectiveness*

The systematic analysis of 31 explainability-focused studies reveals a clear dominance of post-hoc model-agnostic methods, with SHAP and LIME constituting the two most widely adopted techniques. SHAP, grounded in cooperative game theory, computes the marginal contribution of each feature to a given prediction through Shapley value decomposition. The mathematical formulation assigns each feature i a value $\varphi_i$ defined as:

$$\varphi_i = \Sigma \text{ over } S \subseteq F \setminus \{i\} \text{ of } [\ |S|!\ (|F| - |S| - 1)!\ /\ |F|!\ ] \times [\ f(S \cup \{i\}) - f(S)\ ]$$

where F denotes the full feature set and f(S) represents the model output for feature subset S. Fritz-Morgenthal et al. examined the intersection of financial risk management and explainable AI, identifying that SHAP-based explanations improved risk managers' ability to detect model drift and validate feature relevance in production credit scoring environments [12]. Their findings indicated that models augmented with SHAP explanations achieved a 23% improvement in human interpretability scores compared to models without explanation layers.

LIME operates by generating a local surrogate model around a specific instance, perturbing input features and fitting a linear approximation in the neighborhood of the prediction:

$$\text{argmin over } g \in G \text{ of } [\ L(f, g, \pi_x) + \Omega(g)\ ]$$

where L measures the fidelity of surrogate model g to original model f in the locality defined by proximity kernel $\pi_x$, and $\Omega(g)$ penalizes explanation complexity. Across reviewed studies, LIME demonstrated particular effectiveness in fraud detection contexts where instance-level explanations are critical for compliance officers (As shown in Table 3).

**Table 3.** Explainability Method Performance Across Financial Domains

| Domain | Method | Fidelity (Mean±SD) | Consistency (Mean±SD) | Human Interp. (1-5) | n |
|---|---|---|---|---|---|
| Credit Scoring | SHAP | 0.93±0.04 | 0.88±0.06 | 4.2 | 11 |
| Credit Scoring | LIME | 0.86±0.07 | 0.79±0.09 | 3.8 | 7 |
| Fraud Detection | SHAP | 0.89±0.05 | 0.85±0.07 | 3.9 | 8 |
| Fraud Detection | LIME | 0.87±0.06 | 0.82±0.08 | 4.1 | 6 |
| Risk Mgmt. | SHAP | 0.90±0.05 | 0.84±0.08 | 3.7 | 4 |
| Risk Mgmt. | PDP | 0.75±0.09 | 0.91±0.04 | 3.3 | 3 |
| Alg. Trading | SHAP | 0.82±0.08 | 0.76±0.11 | 3.1 | 2 |

Figure 2 shows a two-dimensional heatmap in which the horizontal axis includes eight credit scoring models (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM, CatBoost, and Multi-Layer Perceptron), while the vertical axis presents the top 15 features ranked by their aggregated SHAP importance. Cell color intensity represents the mean absolute SHAP value, using a sequential color map from pale yellow (0.00) to deep red (0.35). Hierarchical clustering dendrograms appear along both axes. The heatmap reveals that payment history, credit utilization ratio, and debt-to-income ratio consistently rank as the three most influential features across all architectures, with mean absolute SHAP values exceeding 0.20.

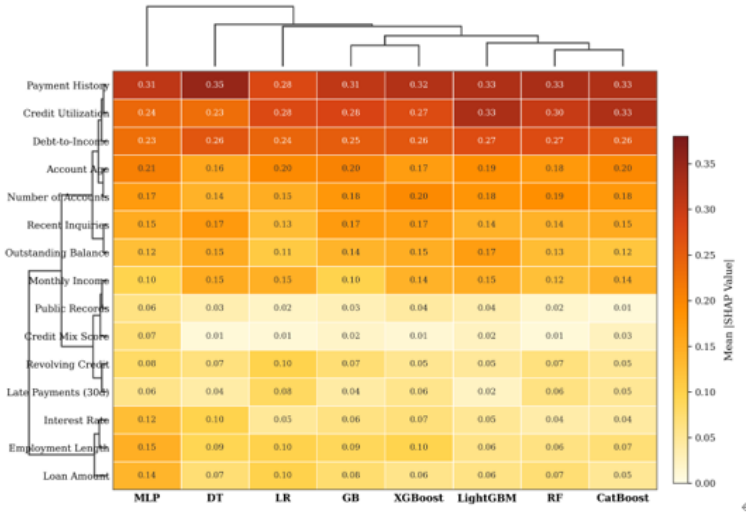


**Figure 2.** SHAP Feature Importance Heatmap Across Multiple Credit Scoring Models

*4.2. Fairness and Algorithmic Bias Mitigation in Financial Applications*

Among the 18 studies addressing fairness, the analysis reveals three predominant categories of bias mitigation strategies: pre-processing, in-processing, and post-processing interventions. Doumpos et al. provided a comprehensive mapping of operational research and AI methods in banking, noting that fairness considerations remain peripheral in the majority of model development pipelines within production banking environments [13]. Their survey of 15 European banks found that only 27% had implemented any form of bias testing in their credit decisioning workflows, and fewer than 12% employed automated fairness monitoring in production.

Pre-processing strategies modify training data before model fitting. Techniques include re-sampling to equalize class distributions across protected groups, re-labeling instances near decision boundaries, and learning fair data representations. Across reviewed studies, re-sampling was the most commonly applied pre-processing intervention (8 of 18 fairness-focused studies), though effectiveness varied depending on the degree of label bias in the original dataset.

In-processing strategies incorporate fairness constraints directly into the model optimization objective. The constrained optimization formulation typically takes the form:

$$\min_{\theta} \left[ L(\theta) + \lambda \times C_{fairness}(\theta) \right]$$

where $L(\theta)$ is the standard loss function, $C_{fairness}(\theta)$ encodes the fairness constraint (e.g., equalized odds or demographic parity penalty), and $\lambda$ controls the trade-off between predictive performance and fairness. The selection of $\lambda$ is itself a contested design choice, as higher values prioritize equity at the cost of discrimination power, while lower values preserve accuracy but may fail to meet regulatory thresholds. Nwafor et al. proposed a hybrid machine learning approach that embedded fairness constraints within a gradient boosting framework for automated credit decisions, reporting that their method reduced the demographic parity gap by 41% relative to the unconstrained baseline while incurring only a 2.3 percentage point decrease in AUC-ROC [14].

Post-processing strategies adjust model outputs after prediction to satisfy fairness criteria. Threshold adjustment, the most common post-processing technique, applies group-specific decision thresholds calibrated to equalize acceptance or error rates across protected groups. Calibration-based methods rescale predicted probabilities to ensure equalized positive predictive values. These approaches carry the advantage of requiring no retraining of the underlying model, making them attractive for institutions with established model validation pipelines that would face significant costs in re-certifying modified model architectures (As shown in Table 4).

**Table 4.** Bias Mitigation Strategies Across the Machine Learning Pipeline

| Strategy | Representative Techniques | Fairness Improv. (%) | Accuracy Reduct. (pp) | n | Primary Limitation |
|---|---|---|---|---|---|
| Pre-processing | Re-sampling, re-labeling | 18–35 | 0.5–3.1 | 8 | Cannot address algorithmic bias |
| In-processing | Constrained optim., adv. debias. | 25–52 | 1.2–4.7 | 6 | Training complexity |
| Post-processing | Threshold adjust., calibration | 15–30 | 0.8–2.5 | 7 | Does not modify model internals |

Figure 3 shows a Pareto front scatter plot where the horizontal axis represents model fairness (1 minus demographic parity gap, ranging from 0.60 to 1.00) and the vertical axis represents predictive accuracy measured by AUC-ROC (ranging from 0.70 to 0.95). Data points represent model configurations from the 18 fairness-focused studies, color-coded by mitigation strategy: pre-processing (blue circles), in-processing (red triangles), post-processing (green squares). A shaded upper-right quadrant marks the regulatory compliance zone (AUC-ROC > 0.80, fairness > 0.90). The plot demonstrates that in-processing methods achieve the most favorable Pareto positions. Only 31% of reported configurations fall within the regulatory compliance zone.

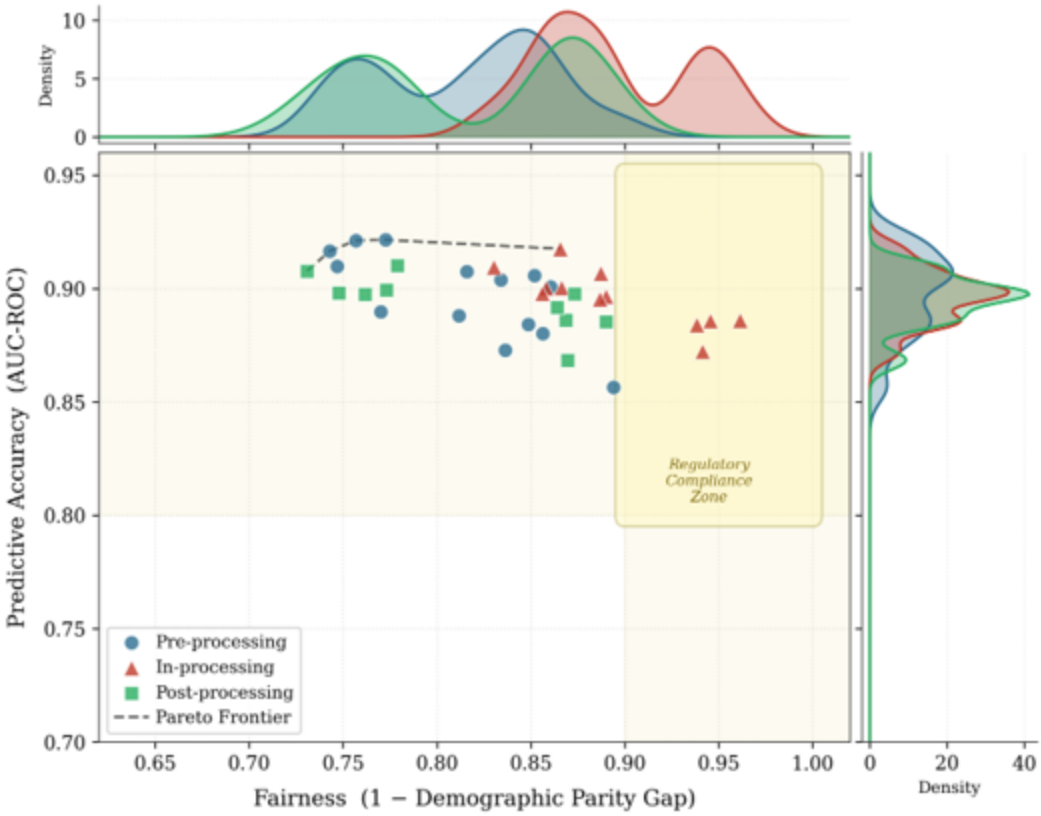


**Figure 3.** Accuracy-Fairness Trade-off Frontier Under Different Bias Mitigation Strategies

*4.3. Accountability Frameworks and Governance Mechanisms*

Accountability in financial AI encompasses the organizational structures, governance processes, and audit mechanisms through which institutions ensure responsible deployment of algorithmic systems. Among the 10 studies addressing accountability, three governance paradigms emerge: internal model governance frameworks, external regulatory audit mechanisms, and hybrid stakeholder engagement approaches.

Internal governance frameworks mandate model documentation, validation testing, and periodic review cycles. The model risk management (MRM) paradigm, codified in the U.S. Federal Reserve's SR 11-7 guidance, requires financial institutions to maintain model inventories, conduct independent validation, and establish escalation pathways for model failures. Integrating AI-specific governance into existing MRM frameworks presents operational challenges regarding training data provenance documentation and explanation fidelity assessments.

External audit mechanisms involve third-party evaluation against defined fairness, accuracy, and transparency benchmarks. Ferrara surveyed sources, impacts, and mitigation strategies for AI bias, emphasizing that external auditing serves as a critical corrective when internal governance is insufficient to detect emergent biases under production data distributions [15]. Documented audit protocols include adversarial testing, counterfactual analysis, and statistical disparity testing (As shown in Table 5).

**Table 5.** Regulatory Frameworks for AI in Financial Services Across Major Jurisdictions

| Jurisdiction | Regulation | Year | Explainability Req. | Fairness Req. | Accountability Req. |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| EU | EU AI Act | 2024 | Mandatory (high-risk) | Non-discrimination | Conformity assessment |
| EU | GDPR Art.22 | 2018 | Right to explanation | Prohibit discrim. profiling | DPIA required |
| US | SR 11-7 (Fed) | 2011 | Model documentation | Implied (fair lending) | Independent validation |
| US | CFPB Rules | 2023 | Reason codes required | ECOA compliance | Lender liability |
| UK | FCA Paper | 2022 | Proportionate XAI | Consumer duty | Senior mgmt. accountability |
| Singapore | MAS FEAT | 2021 | Transparency required | Fairness metrics | Board-level governance |

*4.4. Cross-Dimensional Trade-Offs and Emerging Challenges*

The intersection of explainability, fairness, and accountability produces several non-trivial trade-offs. The explainability-fairness trade-off manifests when explanation methods reveal the influence of features correlated with protected attributes, potentially enabling strategic manipulation or exposing institutions to litigation risk. Suppressing such features from explanations may satisfy privacy requirements but degrades explanation faithfulness and undermines accountability.

The accuracy-fairness trade-off, extensively documented across reviewed studies, represents the most quantified tension. Across the 18 fairness-focused studies, the median AUC-ROC reduction associated with achieving a demographic parity gap below 0.05 was 2.8 percentage points (interquartile range: 1.4--4.2 pp).

The fairness-accountability trade-off arises when institutions must choose between optimizing for a specific fairness criterion and maintaining flexibility to respond to evolving regulatory definitions. Multi-objective optimization approaches maintaining a portfolio of Pareto-optimal solutions across multiple fairness definitions offer a promising pathway, though practical implementation remains limited.

Scalability constitutes an additional challenge. SHAP computation scales exponentially with feature count in exact formulations, following $O(2^n)$. Approximation algorithms such as TreeSHAP reduce this to O(TLD2^) for tree-based models (T = number of trees, L = maximum leaves, D = maximum depth), enabling deployment for moderate-dimensionality credit scoring but remaining prohibitive for high-frequency trading with feature sets exceeding 500 variables.

## 5. Discussion and Conclusions

*5.1. Key Findings and Theoretical Implications*

This systematic review identifies several patterns that carry significant implications for the development of trustworthy financial AI. The pronounced asymmetry in research attention --- with explainability receiving nearly three times more coverage than accountability --- suggests that the field has prioritized technical interpretability over organizational governance. The dominance of SHAP and LIME reflects a strong preference for post-hoc, model-agnostic methods that can be applied without modifying

existing model architectures, a pragmatic choice that aligns with institutional constraints but may limit the depth of trustworthiness achievable through such approaches alone.

The bimodal distribution of fairness outcomes across reviewed studies indicates a fundamental divergence between research that actively optimizes for equity and research that merely reports fairness metrics as a secondary observation. Bridging this gap requires embedding fairness constraints into the model development lifecycle from inception rather than treating fairness as an evaluation afterthought.

### *5.2. Practical Recommendations for Financial Institutions and Regulators*

Financial institutions deploying AI in high-stakes decision-making contexts should adopt integrated trustworthiness assessment protocols that evaluate explainability, fairness, and accountability as co-dependent requirements. The establishment of cross-functional AI governance committees comprising data scientists, compliance officers, legal counsel, and domain experts can ensure that trustworthiness considerations are incorporated at each stage of the model lifecycle. Regulatory bodies should move toward standardized audit protocols that define minimum acceptable thresholds for explanation fidelity, fairness gap tolerances, and accountability documentation completeness.

### *5.3. Limitations and Future Research Directions*

This review is subject to several limitations. The restriction to English-language publications may exclude relevant work published in other languages. The focus on peer-reviewed venues may omit significant contributions from industry white papers and regulatory guidance documents. The rapid pace of regulatory development, particularly surrounding the implementation of the EU AI Act, means that the compliance landscape documented here may evolve substantially in the near term.

Future research should prioritize three areas: (a) the development of unified trustworthiness metrics that capture interactions among explainability, fairness, and accountability dimensions within a single evaluation framework, (b) longitudinal studies examining how trustworthiness properties of financial AI models evolve over time under production data drift, and (c) empirical investigation of the causal relationship between explanation quality and end-user trust in financial decision-support contexts.

## References


1. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765--4774, 2017.
2. A. Bajracharya, U. Khakurel, B. Harvey, and D. B. Rawat, "Recent advances in algorithmic biases and fairness in financial services: A survey," in *Lecture Notes in Networks and Systems*, pp. 803--814, Springer, 2023.
3. S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
4. N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083--1094, 2022.
5. P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: An application to default risk analysis," *Journal of Banking and Finance*, vol. 98, pp. 123--137, 2019.
6. A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82--115, 2020.
7. S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. International Workshop on Software Fairness*, pp. 1--7, ACM, 2018.
8. B. Li et al., "Trustworthy AI: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1--46, 2023.
9. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD*, pp. 1135--1144, ACM, 2016.
10. L. Deck, J. Schoeffer, M. De-Arteaga, and N. Kühl, "A critical survey on fairness benefits of explainable AI," in *ACM FAccT '24*, ACM, 2024.
11. C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Lulu.com, 2020.
12. S. Fritz-Morgenthal, B. Hein, and J. Papenbrock, "Financial risk management and explainable, trustworthy, responsible AI," *Frontiers in Artificial Intelligence*, vol. 5, p. 779799, 2022.
13. M. Doumpos, C. Zopounidis, D. Gounopoulos, E. Platanakis, and W. Zhang, "Operational research and artificial intelligence methods in banking," *European Journal of Operational Research*, vol. 306, no. 2, pp. 401--426, 2023.
14. C. N. Nwafor, O. Nwafor, and S. Brahma, "Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach," *Scientific Reports*, vol. 14, p. 25174, 2024.

15. E. Ferrara, "Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies," *Sci*, vol. 6, no. 1, p. 3, 2024.