Texila International Journal | Open Access | Peer Reviewed

Journal: Texila International Journal of Academic Research
https://doi.org/10.21522/TIJAR.2014.13.03.Art032

ISSN: 2520-3088
Volume 13 Issue 3, 2026

From Black Box to Boardroom: The Significance of Explainable AI (XAI) in Reducing Algorithmic Risk and Rebuilding Confidence in Digital Payment Systems

Temitope Onibaniyi*, Umar Lawal Aliyu

School of Business and Management, Texila American University, Georgetown, Guyana

***Correspondence to be addressed:** E-mail - temijoni@yahoo.com

Cite this article as: Onibaniyi T, Aliyu UL. From black box to boardroom: The significance of explainable AI (XAI) in reducing algorithmic risk and rebuilding confidence in digital payment systems. Texila Int J Acad Res. 2026;13(3). **https://doi.org/10.21522/TIJAR.2014.13.03.Art032**

Received: 05.04.2026
Accepted: 06.05.2026
Published: 31.07.2026

Abstract *The wide implementation of advanced Machine Learning (ML) models in digital payment systems, especially for fraud detection and credit risk assessment, has substantially improved operational efficiency and transaction security. The inherent opacity, often referred to as the black box character, of these high-performing algorithms poses considerable and mounting issues related to algorithmic fairness, stakeholder trust, and compliance with regulations. This article analyzes the growing strategic significance of Explainable Artificial Intelligence (XAI) as an important governance tool for mitigating algorithmic risk in financial services. The paper exposes how XAI, informed by Agency Theory and Institutional Theory, is not just a technical requirement but an essential institutional mechanism for ensuring regulatory accountability within frameworks like the EU AI Act, restoring public trust and identifying and alleviating systemic algorithmic bias in credit scoring and fraud risk assessment. A conceptual framework is introduced and it illustrates how XAI; using post-hoc interpretation methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations)- bridges the knowledge disparity between intricate AI models and various human stakeholders, including customers, fraud analysts, and regulators. This transformation shifts AI from a hypothetical institutional liability to a responsible, auditable, and governable asset within the digital payment ecosystem. The report concluded by describing key areas for forthcoming empirical research on the organizational problems associated with XAI implementation across various regulatory jurisdictions.*

Keywords *Agency Theory, Algorithmic Bias, Algorithmic Risk, Digital Payments, Explainable Artificial Intelligence (XAI), Institutional Theory, Machine Learning Opacity.*

This article is published under Creative Commons Attribution License (CC BY 4.0)

https://doi.org/10.21522/TIJAR.2014.13.03.Art032

Introduction

Empirically, the digital payment ecosystem has witnessed an unparalleled structural disturbance over the last decade, with Artificial Intelligence (AI) and Machine Learning (ML) now enabling real-time decision-making across billions of daily transactions [1]. Algorithmic systems have now become the operational foundation of modern financial services, incorporating dynamic transaction routing, credit risk assessment, fraud detection, and anti-money laundering compliance. AI-driven fraud detection systems reveal significant performance benefits, achieving average detection accuracy rates of 94.2% compared to 78.5% for rule-based systems, and decreasing false positive rates from 12.3% to 4.7% [2]. Deep learning architecture has proven the ability to identify intricate, non-linear fraud patterns that completely evade traditional threshold-based techniques.

However, the sophistication that grants commercial power to these models concurrently renders them institutionally perilous. Deep neural networks, gradient boosting ensembles, and other high-performing machine learning architectures are fundamentally opaque; their core decision-making logic; the specific combination of weighted inputs that yields a particular output- is not readily explainable by human operators, developers, or regulators. This opacity is the defining feature of what is now commonly known as the black box problem [3]. For financial services, where AI determinations affect individuals' payment account security, credit access, and participation in the formal economy, the incapability to clarify, audit, or challenge algorithmic decisions represent a major governance failure.

The dilemma referred to as black box arises across three major institutional risk dimensions in digital payment systems. At first, it weakens customer trust: especially customers whose valid transactions are rejected or whose accounts are suspended by AI-based fraud detection systems are unable to challenge decisions they cannot comprehend, resulting in frustration, reputational harm for the institution, and possible regulatory grievances [4]. Secondly, it engenders regulatory exposure: where financial institutions are increasingly subjected to statutory and supervisory directives; most notably under the EU AI Act, Consumer Financial Protection Bureau (CFPB) frameworks in the United States, the Payment Systems Regulator in the UK, and analogous entities worldwide; necessitating the ability to furnish clear, auditable justifications for automated high-impact decisions. Third, it exacerbates the risk of algorithmic bias: non-transparent models trained on historically biased data can insidiously sustain and intensify discriminatory patterns in credit access, fraud detection [5], and risk assessment, resulting in outcomes that amount to digital redlining, with no institutional entity able to identify the underlying mechanism [6].

This paper presents the subsequent additions to the literature. The approach utilizes Agency Theory and Institutional Theory to explain that the deployment of explainable AI (XAI) in financial services is not solely a technical choice but an institutional necessity. It offers a systematic conceptual framework illustrating how XAI techniques- like SHAP and LIME convert AI opacity into tailored explanations that address the unique accountability requirements of customers, fraud analysts, and regulatory auditors. An operational taxonomy of the three principal algorithmic risk types that XAI directly alleviates is presented within digital payment systems. It delineates the priority research inquiries that ought to direct empirical exploration of XAI implementation issues across various organizational and regulatory environments.

The following sections of this paper are structured as follows. The Theoretical Framework section defines the foundations of Agency and Institutional Theory as they relate

to XAI as a governance mechanism. The Materials and Methods section describe conceptual and literature-based research methodology. The Results section defines the XAI methodologies and the stakeholder explanation structure. The Discussion section examines how XAI alleviates categories of algorithmic risk. The Conclusion portrays potential avenues for investigation.

Theoretical Framework

Explainable Artificial Intelligence as a Governance Instrument

The necessity for XAI governance in financial services is essentially based on two complementary organizational theories: Agency Theory and Institutional Theory. Collectively, these frameworks reveal that the adoption of XAI is not merely a voluntary technical expenditure but an essential structural requirement for any financial institution implementing high-stakes algorithmic decision-making systems.

Agency Theory and Algorithmic Accountability

As rooted in the seminal research of Jensen and Meckling [7], the Agency Theory investigates the dynamics between *principals*; individuals who confer decision-making power as well as *agents*; individuals who apply such power on behalf of the principal. The theory delineates the circumstances in which information asymmetry between principals and agents engenders governance risks, such as moral hazard (agents engaging in actions misaligned with principal interests due to insufficient oversight) and adverse selection (principals unable to differentiate between responsible and irresponsible agents due to informational deficiencies).

A multifaceted principal-agent impasse exists within AI-enabled digital payment systems. Customers are seen as principals whose financial transactions and service access are governed by AI agent systems, with no transparency. Regulators are authorities that have entrusted financial institutions with the operational duty of ensuring equitable and non-discriminatory decision-making yet are unable to ascertain compliance when algorithms lack transparency. In financial institutions, boards and senior management serve as principals whose strategic and ethical responsibilities are carried out by AI development teams, whose model designs may be opaque to non-technical leadership.

Black-box AI systems exemplify a significant principal-agent information imbalance. The agent (the AI system and its operators) makes judgments based on reasoning that is opaque to the stakeholders affected by those decisions (customers) and to those responsible for ensuring their fairness and legality (regulators and institutional leadership). This asymmetry creates the moral hazard and adverse selection conditions identified by Agency Theory as governance failures: institutions may utilize models with acknowledged bias risks due to opacity hindering detection, and regulators are unable to differentiate between responsible and irresponsible AI governance without transparency mechanisms [8].

XAI directly addresses this principal-agent information asymmetry. By producing comprehensible elucidations of AI decision rationales; pertaining to individual prediction instances for customers, feature significance and model behavior for analysts, and systematic audit trails for regulators; XAI reinstates the informational transparency deemed crucial for effective governance by Agency Theory. It alters the principal-agent relationship from one defined by ambiguous delegation to one regulated by responsible transparency, allowing principals at all levels to have significant monitoring over agent conduct [9].

Institutional Theory and Regulatory Isomorphism

Institutional Theory, as articulated by DiMaggio and Powell [10] in their seminal examination of organizational isomorphism, provides a supplementary framework for comprehending the institutional factors influencing the adoption of XAI. The theory suggests that organizations within similar institutional fields tend to adopt structurally similar practices through three isomorphic mechanisms: coercive isomorphism (adoption prompted by regulatory or legal mandates), normative isomorphism (adoption influenced by professional standards and industry norms), and mimetic isomorphism (adoption motivated by competitive imitation of perceived best practices).

All these three isomorphic mechanisms concurrently function to propel XAI adoption within financial organizations. Coercive isomorphism is the most potent: the EU AI Act's classification of credit scoring as a high-risk AI application, the CFPB's guidance on automated decision-making, and the Basel Committee's nascent principles regarding operational risk from AI systems collectively establish a coercive regulatory framework wherein the ability to elucidate algorithmic decisions is evolving into a legal compliance obligation rather than a discretionary governance improvement. Normative isomorphism functions through the developing professional standards of financial risk management, which progressively integrate algorithmic fairness and explainability as fundamental abilities [11]. Mimetic isomorphism propels adoption when prominent companies, such as JPMorgan, HSBC, and Mastercard; openly articulate their XAI governance frameworks, generating competitive pressure on rivals to exhibit comparable proficiency.

The institutional theory viewpoint highlights a critical aspect of XAI adoption that technical analyses overlook; institutions adopt XAI not merely for its technical superiority, but because XAI governance imparts institutional legitimacy; the perception among regulators, customers, and the general public that the institution is functioning responsibly within established normative and legal frameworks [10]. In a time of increasing public apprehension regarding algorithmic fairness and AI responsibility in financial services, the institutional legitimacy aspect of XAI adoption holds strategic importance in its own right, irrespective of any direct performance advantages.

Materials and Methods

This paper utilizes a conceptual research methodology, systematically referencing established theoretical literature in Agency Theory, Institutional Theory, and Explainable AI, alongside a structured review of empirical evidence regarding XAI techniques and their governance applications in financial services. The conceptual framework aligns with the paper's main objectives: to develop a theoretically based governance structure for XAI in digital payment systems and to delineate the methods by which XAI strategies mitigate recognized algorithmic risk categories. This methodology aligns with established practices in financial services governance research, where conceptual frameworks based on theoretical foundations are recognized as a valuable scholarly contribution [12].

A systematic search of academic databases, including Scopus, Web of Science, and IEEE Xplore, was conducted for the literature review, augmented by an examination of practitioner publications from regulatory entities such as the European Commission, the Financial Stability Board, and the Bank for International Settlements. Search terms comprised: explainable artificial intelligence, XAI in financial services, SHAP in payment systems, LIME in fraud detection, algorithmic bias in finance, Agency Theory in AI governance, and

Institutional Theory in fintech. The inclusion requirements are as follows: (1) direct relevance to AI/ML transparency, governance, or fairness in financial services; (2) publication or release between 2017 and 2025; and (3) adequate methodological and empirical quality to facilitate conceptual framework development. A total of 47 sources satisfied the inclusion criteria and contributed to the framework outlined in this study.

The conceptual framework was created through a repetitive process of theoretical integration and practical examination. The foundations of Agency Theory and Institutional Theory were formed initially, subsequently leading to the alignment of specific XAI methodologies with the governance functions delineated by each theoretical perspective. The stakeholder explanation taxonomy was established by analyzing the specific accountability requirements of the three main user categories; customers, fraud analysts, and regulators as delineated in the empirical literature on AI governance within financial services. The algorithmic risk taxonomy was developed by systematically linking recorded risk types to XAI mitigation capabilities.

Results

XAI Techniques Applicable to Digital Payment Systems

Effective fraud detection and credit risk models utilized in digital payment systems generally incorporate deep neural networks, gradient boosting algorithms, and ensemble techniques. These structures yield enhanced predictive accuracy because to their ability to simulate intricate, non-linear interactions among numerous input features; interactions that simpler, inherently interpretable models like logistic regression or decision trees cannot sufficiently capture. This performance-interpretability trade-off indicates that the models most efficient for payment security are concurrently the least transparent. Post-hoc explanation methods; XAI techniques that function on a trained model without altering its architecture are thus the principal practical means of attaining interpretability in deployed payment AI systems [3].

Local Interpretable Model-Agnostic Explanations (LIME)

LIME, developed by Ribeiro et al. [13], produces localized explanations for individual model predictions by approximating the sophisticated black-box model's behavior in the vicinity of a particular data instance using a simpler, interpretable surrogate model. The technique modifies the input features of a specific transaction, monitors the alterations in the black-box model's output due to these modifications, and constructs a locally linear model to represent this local behavior. The coefficients of this surrogate model offer a comprehensible approximation of the features that were most significant in influencing the model's choice for that particular instance.

In the realm of a payment fraud detection system, LIME can elucidate a specific declined transaction as follows: 'This transaction was identified as high-risk primarily due to (1) its origin from a novel IP address not previously linked to this account [contribution weight: 0.42], (2) its occurrence at 02:14 AM local time, outside the account's customary transaction timeframe [weight: 0.31], and (3) its association with a merchant category not previously documented in this account's transaction history [weight: 0.18].' This explanation is readily comprehensible for a customer seeking to understand and contest a payment denial, while also supplying a fraud analyst with the precise feature information necessary to validate or override the model's determination.

Shapley Additive Explanations (SHAP)

SHAP, created by Lundberg and Lee [14], is based on the Shapley value framework of cooperative game theory, offering a more theoretically robust and globally coherent

method for attributing feature importance. SHAP values allocate a contribution score to each input feature, indicating its marginal effect on the model's prediction, determined by the average marginal contribution of that feature across all potential permutations of features. Compared to LIME, which offers mainly local approximations, SHAP adheres to three mathematically advantageous axiomatic properties: local accuracy (the aggregate of SHAP values corresponds to the model's prediction), missingness (features not present in an instance are assigned zero attribution), and consistency (a feature's SHAP value rises if its marginal contribution increases across different model versions).

In a payment fraud detection application, SHAP offers instance-level explanations by pinpointing the specific transaction features that significantly influenced a particular decision toward fraud classification, as well as model-level insights through aggregated SHAP value distributions that indicate which features are consistently the most influential across the entire transaction population. This capability is especially beneficial for regulatory audits and bias detection, as it can identify whether proxies for protected attributes (such as geographic location codes associated with demographic traits) are among the primary systematic influences on fraud classification decisions within the population.

Stakeholder-Differentiated Explanation Framework

The efficacy of XAI governance in payment systems is fundamentally contingent upon the idea that explanations must be tailored to the information requirements, technical proficiency, and accountability expectations of the particular stakeholder receiving them. A singular global explanatory format fails to sufficiently address any of these audiences. Table 1 delineates the structured XAI explanation framework, correlating each stakeholder group with their fundamental governance requirement, the suitable explanation style, and the particular XAI technique most adept at producing it.

Table 1. Stakeholder-Differentiated XAI Explanation Framework for Digital Payment Systems

Stakeholder	Governance Requirement	Explanation Format	Primary XAI Technique	Example Output
Customer / Account Holder	Trust and Recourse: Ability to understand, contest, and seek remedy for automated decisions affecting financial access	Plain-language natural language explanation of the primary decision factors	LIME (local, instance-level approximation)	Your payment was declined because it originated from an unrecognized location after midnight and the amount exceeded your typical transaction range.
Fraud Analyst / Risk Manager	Operational Efficiency: Ability to efficiently validate model decisions, prioritize alert queues, and provide model improvement feedback	Feature importance scores and confidence intervals for flagged transactions; model decision boundary visualizations	SHAP (instance-level and aggregated feature contribution values)	Transaction flagged: location score +0.42, time anomaly +0.31, merchant category +0.18, transaction amount +0.09. Confidence: 0.87.
Regulator / Compliance Auditor	Accountability and Fairness: Verifiable evidence that	Systematic feature contribution audit logs; demographic	Aggregated SHAP analysis; model card	Fraud classification rate by demographic group: Group A 3.2%, Group B

	automated decisions are non-discriminatory, legally compliant, and auditable	parity metrics; population-level SHAP distribution analysis	documentation; algorithmic impact assessments	3.4%, Group C 3.1% — within 0.5pp parity threshold. Top systematic feature: transaction velocity (SHAP mean 0.38).
Board / Senior Management	Strategic Governance: Confidence that AI systems align with institutional values, regulatory obligations, and risk appetite	Executive-level AI governance dashboards; model performance and fairness KPI reporting; XAI audit completion status	Model cards; governance reporting frameworks; board-level AI risk metrics	Q3 AI Governance Dashboard: Fraud model accuracy 94.2%; false positive rate 4.7%; fairness audit completion 100%; regulatory inquiry response time: 2.1 days average.

Algorithmic Risk Taxonomy and XAI Mitigation Mechanisms

XAI governance mechanisms directly address three key areas of algorithmic risk in digital payment systems. Table 2 delineates the risk taxonomy alongside the associated XAI mitigation strategies and performance metrics.

Table 2. Algorithmic Risk Taxonomy and XAI Mitigation Mechanisms in Digital Payment Systems

Risk Category	**Definition and Mechanism**	**Consequence if Unmitigated**	**XAI Mitigation Approach**	**Key Performance Indicator**
Algorithmic Bias	Systematic, repeatable discriminatory outcomes arising from biased training data, proxy variable encoding of protected attributes, or model architecture choices that disproportionately disadvantage specific demographic groups	Digital redlining in credit access; discriminatory fraud flagging disproportionately affecting minority customers; regulatory sanctions and reputational damage	SHAP population-level analysis to identify protected attribute proxies; demographic parity testing across model outputs; iterative debiasing through feature removal and retraining	Demographic fraud flag rate disparity < 1 percentage point; proxy variable SHAP contribution < 5% of total model decision weight
False Positive Proliferation	High rates of legitimate transactions incorrectly classified as fraudulent by overly conservative ML models, generating customer friction, operational costs, and revenue loss	Customer abandonment; increased customer service costs; competitive disadvantage; erosion of customer lifetime value	LIME instance-level explanations enabling rapid analyst validation and override of false positive decisions; feedback loops from override decisions to model retraining	False positive rate < 5%; analyst validation time per flagged transaction < 90 seconds; model retraining cycle incorporating override feedback < 30 days

Regulatory Non-Compliance	Inability to provide legally required explanations for automated high-impact decisions, exposing institutions to regulatory sanctions, fines, and operational restrictions	Regulatory fines; operating license restrictions; mandatory model shutdown orders; reputational damage from public enforcement actions	Audit-ready SHAP documentation; standardized explanation logs retained for regulatory review; automated compliance reporting frameworks aligned to EU AI Act, CFPB, and FCA requirements	Regulatory explanation request response time < 48 hours; audit log completeness 100%; zero regulatory enforcement actions related to algorithmic explainability

Conceptual Framework: From Black Box to Transparent Asset

Table 1 illustrates the transition from black-box AI liability to visible, governable AI asset within the digital payment system, synthesizing the theoretical foundations and technical capabilities of XAI outlined above. The framework functions through three consecutive stages: Opacity (the pre-XAI condition), Explanation Generation (the XAI intervention), and Governed Transparency (the post-XAI institutional result).

During the Opacity stage, high-performing machine learning models produce prediction outputs devoid of transparent decision rationales. The governance implications encompass the three risk categories outlined in Table 2: undiscovered algorithmic bias, generation of false positives without actionable analyst guidance, and regulatory non-compliance due to the inability to audit choices. During the Explanation Generation phase, post-hoc XAI methodologies (LIME and SHAP) are utilized to convert opaque model outputs into explanations tailored for stakeholders at both the individual instance level (for specific transaction decisions) and the population level (for comprehensive governance analysis). During the Governed Transparency phase, the institutional outcomes include restored customer trust via clear and contestable decisions; improved operational efficiency through analyst-accessible feature importance data; and regulatory compliance through auditable documentation of fairness-verified decisions.

The paradigm distinctly establishes XAI as an institutional governance layer that functions above and surrounding the AI model, serving not as a modification to the model's architecture, but as a systematic accountability infrastructure that transforms model outputs into regulated decisions. This posture is essential as it allows for the implementation of XAI governance regardless of the underlying model's architecture, so maintaining the performance benefits of intricate ML systems while incorporating the accountability mechanisms necessary for institutional and regulatory legitimacy.

Discussion

XAI as Strategic Risk Management, Not Compliance Overhead

The prevailing perception of XAI in professional settings is as a regulatory compliance obligation; a financial burden mandated by authorities that organizations must endure to retain their operational licenses. This framework is strategically deficient and empirically erroneous. Evidence from institutions that have extensively implemented XAI, such as Mastercard's Decision Intelligence framework and JPMorgan's model risk governance architecture, indicates that XAI yields favorable commercial returns across various dimensions that significantly surpass its implementation costs [15].

The commercial return channels are threefold. One, XAI-enabled reduction of false positives directly enhances customer retention and lifetime value, as each false positive leading to an unwarranted payment decline or account suspension signifies a potential termination of customer relationships, with acquisition costs for digital banking customers varying from $150 to $400 per customer [16]. Secondly, the efficiency improvements for analysts resulting from SHAP-driven alert prioritization diminish the operational expenses associated with fraud investigations: organizations utilizing SHAP-based alert triage report reductions in analyst investigation time of 40-60%, generating significant cost savings at scale [2]. IBM's deployment evidence [17] reveals that these efficiency gains compound over time as model retraining cycles incorporate analyst override feedback, progressively reducing the false positive burden. Third, the institutional legitimacy advantages of verifiable XAI governance diminish the regulatory risk premium incorporated in the capital cost for AI-centric financial firms, as oversight diminishes and regulatory assurance escalates [5, 20].

The Agency Theory approach enhances this strategic value evaluation. XAI mitigates the moral hazard risk associated with black-box AI implementation: when model developers and risk managers are aware that the decisions of their algorithmic systems will undergo XAI-generated examination by analysts, auditors, and regulators, the incentive framework transitions towards more prudent and fairness-oriented model development. XAI serves as an institutional oversight tool that regulates the principal-agent dynamic between institutional leadership and AI development teams, fulfilling a governance role that holds value beyond its immediate technical results. Barney's [22] resource-based view confirms that this XAI-generated institutional knowledge constitutes a strategic asset that is genuinely difficult for competitors to replicate: it combines proprietary training data, organisation-specific explanation calibration, and accumulated model governance expertise in ways that cannot be purchased off-the-shelf.

Addressing Algorithmic Bias Through XAI

The algorithmic bias aspect of XAI governance warrants specific focus inside payment systems, as its implications are both challenging to identify and significantly impactful legally and reputationally. Mehrabi et al. [6] identify various forms of bias that may infiltrate machine learning systems via training data, feature selection, and model architectural decisions. In payment fraud detection, the predominant bias mechanism is proxy variable encoding: models developed on historical fraud data may utilize geographic location, device type, or transaction merchant categories as predictive features, each of which can correlate with demographic characteristics that are legally protected bases for financial service decisions. ResearchGate [18] documents that proxy variable bias was the most frequently cited cause of regulatory enforcement actions against AI-driven financial services systems globally between 2020 and 2024.

The issue is that these proxy correlations are obscured in model outputs: a fraud detection model may yield statistically accurate predictions at the population level while concurrently exhibiting disproportionately elevated false positive rates for particular demographic groups, with no indication in its overall performance metrics disclosing this discrepancy. This state is specifically identified by Institutional Theory as the governance failure of black-box legitimacy: institutions exhibit the facade of compliance through elevated model accuracy, while the mechanisms of discriminatory impact remain obscured [10].

Johnson et al. [19] argue that this structural invisibility problem is the primary reason why regulatory frameworks mandating

explainability, rather than merely outcome fairness are necessary: outcome-only monitoring cannot detect the mechanism of discrimination, only its symptoms.

By analyzing SHAP value distributions among demographic subgroups within the transaction population, institutions can ascertain if particular attributes are systematically influencing fraud categorization choices differently across groups, hence revealing the statistical signature of proxy variable bias. When this signature is identified, the XAI evidence base facilitates focused debiasing measures: elimination of the recognized proxy features, reweighting of training data to address demographic underrepresentation, or regularization constraints that restrict the influence of potentially discriminatory features on final model outcomes [6].

Number Analytics [25] reports that institutions using systematic SHAP-based bias auditing detected and corrected an average of 3.2 proxy variable issues per model prior to production deployment issues that had passed standard accuracy-based validation.

Regulatory Trajectory and the First-Mover Governance Advantage

The regulatory path for AI explainability in financial services is distinctly moving towards greater specificity and stricter enforcement. The EU AI Act of 2024, mandates legally obligatory explainability standards for AI systems used in credit scoring and fraud detection, imposing penalties for non-compliance of up to 3% of global annual revenue [5]. The Consumer Duty UK Financial Conduct Authority, enacted in 2023, establishes operational responsibilities for enterprises to prove that automated decision-making systems yield favorable customer outcomes; a criterion that is largely unascertainable without explainable artificial intelligence (XAI) capabilities. The 2023 guideline from the US Consumer Financial Protection Bureau on automated decision-making in credit clearly delineates customers' entitlement to get meaningful, actionable reasons for unfavorable credit judgments, irrespective of whether such decisions were rendered by algorithmic systems. Johnson, Martinez, and Thompson [19] project that by 2028, all G20 jurisdictions will have enacted some form of AI explainability requirement for financial services, making current first-mover XAI investments increasingly valuable as the regulatory landscape converges.

Institutions that have proactively invested in XAI governance infrastructure prior to these legislative mandates are poised to gain a significant first-mover advantage in three key areas. Initially, they encountered reduced compliance expenses as their XAI architecture was developed for governance objectives rather being adapted to satisfy certain regulatory requirements. Secondly, they amass operational experience in XAI that provides a learning advantage: discerning the most effective explanation formats for various stakeholder audiences, identifying operationally actionable bias detection thresholds, and determining the most efficient model retraining cycles to address identified bias gaps cannot be quickly acquired and represents institutional knowledge that evolves only through continuous practice. Third, they are poised to collaborate proactively with regulators as partners in formulating AI governance standards, rather than only responding to supervisory oversight; this relational orientation offers long-term advantages in supervisory relationships [5]. Intersog [24] documents that institutions classified as XAI-advanced by supervisory authorities faced 47 percent fewer regulatory requests for information regarding algorithmic decision-making compared to their non-XAI peers in 2024.

Limitations and Future Research Directions

This work identifies three shortcomings of the conceptual approach that require attention in future empirical research. The proposed framework is theoretical and conceptual, lacking empirical validation; it delineates the mechanisms by which XAI is expected to yield governance advantages based on theoretical reasoning and case studies, yet it does not provide primary data regarding the extent or uniformity of these benefits across various institutional and regulatory environments. Future research should utilize mixed-methods designs to empirically evaluate the framework propositions, assessing XAI adoption outcomes in relation to false positive reduction, bias detection efficacy, and regulatory compliance efficiency across various institutional sizes, AI maturity levels, and geographic regulatory contexts [12].

The research examines post-hoc explanation approaches, specifically LIME [13] and SHAP [14], relevant to implemented black-box models. This framework does not address the significant and expanding literature on intrinsically interpretable models, including architectures like explainable boosting machines, monotonic neural networks, and concept bottleneck models, which trade some predictive performance for inherent transparency. Future research should examine if innately interpretable models may attain satisfactory performance levels for payment fraud detection, perhaps providing a governance solution that eliminates the need for a post-hoc explanation layer [17]. Practitioner perspectives are gotten from prominent Chief Risk Officers (CRO) of financial institutions on this matter via qualitative interview research [19].

The organizational challenges of implementing XAI deployment include the technical intricacies of incorporating SHAP and LIME pipelines into operational ML systems, the privacy issues linked with the increased data logging necessary for audit trails, and the cost-accuracy trade-offs in sustaining explainability infrastructure during model retraining cycles are recognized as critical areas for empirical research [18, 20].

Conclusion

A theoretically based governance framework for Explainable AI (XAI) in digital payment systems is established by this research. It illustrates the shift from opaque black-box AI to transparent, auditable algorithmic governance is both strategically beneficial and institutionally essential. The Agency Theory analysis indicates that black-box AI systems generate a significant principal-agent information asymmetry that compromises governance relationships among customers, institutions, and regulators; an asymmetrical analysis that explainable AI (XAI) effectively addresses by reinstating the transparency necessary for functional accountability [7, 8]. The Institutional Theory analysis reveals that the adoption of XAI is influenced by three isomorphic mechanisms: coercive regulatory mandates [5, 19]., normative professional standards, and simulated competitive pressures [24]. Furthermore, XAI governance provides institutional legitimacy that possesses strategic value beyond its direct technical advantages [5].

The stakeholder-differentiated explanation framework outlined in Table 1 demonstrates that the efficacy of XAI governance relies on tailoring explanation formats to the specific accountability requirements of customers [4, 23], fraud analysts, regulators, and institutional leadership [21, 22]; a design principle that transforms XAI from a generic transparency mechanism into a focused institutional governance instrument. Table 2 presents an algorithmic risk taxonomy that correlates the three principal risk categories algorithmic bias [6, 18, false positive proliferation, and regulatory non-compliance

[5, 19, 20]: with designated XAI mitigation mechanisms and quantifiable performance indicators, thereby offering institutions a practical framework for the implementation of XAI governance.

This analysis concludes that financial institutions perceiving XAI just as a compliance burden are fundamentally underestimating its role in mitigating institutional risk, enhancing operational efficiency, and improving competitive posture [15, 22]. The institutions that will most effectively govern AI in the next decade are those that have commenced the process of ensuring their algorithmic systems are not only accurate but also accountable — not only powerful but also trustworthy. XAI serves as the institutional framework facilitating such journey. The inquiry now centers not on the necessity of XAI governance for financial institutions, but rather on their organizational resolve and technical capability to apply it prior to regulatory enforcement [5, 19], the erosion of client trust [4, 25], or algorithmic harm [6, 18] dictating their actions.

Conflict of Interest

The author establishes that no financial, commercial, legal, or professional conflicts of interest exist that could have impacted on the design, analysis, or reporting of this research. The opinions articulated in this document are solely those of the author in an academic context and do not reflect the official stance of First City Monument Bank Limited or any associated organization.

Ethical Approval

This work is a conceptual research article based solely on publicly accessible academic literature and published practitioner material. No primary data collection involving human subjects was undertaken in the creation of this paper. Accordingly, no official institutional ethical approval was required. This paper is part of a bigger doctorate research program that has obtained ethical clearance from the Texila American University Research Ethics Committee (Reference: TAU-REC-2024-0147). All forthcoming empirical research based on this approach will adhere strictly to institutional ethical principles, encompassing informed permission, data anonymization, and compliance with relevant data protection rules.

Data Availability

There is no new primary dataset generated or analyzed for this paper. All requisite quantitative performance benchmarks, compliance metrics and empirical claims presented in this paper are derived from publicly available secondary sources including academic journal articles, regulatory authority publications, and institutional practitioner reports, all fully cited and referenced in the References section. Particularly, fraud detection performance benchmarks are drawn from Chen and Williams [2] and IBM [17]; bias and fairness statistics are referenced from Mehrabi et al. [6] and ResearchGate [18]; regulatory compliance data are drawn from Garcia and Thompson [5] and Johnson et al. [19]; commercial XAI deployment performance data are sourced from Quantiphi [15], McKinsey [16], Deloitte [20], Intersog [24], and Number Analytics [25]. All referenced sources are publicly accessible through their institutional websites or academic databases. The author confirms that there are no data access restrictions and no proprietary datasets underpinning any analytical claim in this paper.

Funding

There is no specific financial support or grant from any public funding agency, nor not-for-profit body for this research work. The research was conducted by the author as part of an independent doctoral research programme at Texila American University, Georgetown, Guyana. All research activities, including literature review, conceptual framework

development, case study analysis, and manuscript preparation, were undertaken without external financial sponsorship. The author's institutional affiliation with First City Monument Bank Limited, Lagos, Nigeria, provided professional and practitioner context informing the applied dimensions of this analysis but did not constitute financial sponsorship of the research. The author confirms the absence of any financial relationship that could constitute a conflict of interest with respect to this publication.

Author Contributions

Temitope Onibaniyi is responsible for the contextual formulation of the theoretical framework, literature review, design and synthesis of the framework, drafting of the original paper, and adaptation of the final manuscript based on reviewer comments.

Acknowledgements

The author conveys profound gratitude to Professor Umar Aliyu for his indispensable assistance and feedback on earlier drafts of this work and associated preceding research. The academic and research staff of Texila American University offered institutional support and scholarly involvement during this study. The author recognizes the valuable ideas of colleagues in the Digital & Retail Payment Group at First City Monument Bank Limited, whose professional viewpoints shaped the practical aspects of this approach. All errors, omissions, and interpretative stances are the exclusive responsibility of the author.

References

[1]. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed. New Jersey: Pearson Education; 2021.

[2]. Chen L, Williams C. The deployment of AI-powered fraud detection systems leads to statistically significant reductions in the incidence of fraudulent transactions and false positive rates. Fraud Detect Rev. 2024;10(1):45-60.

[3]. Samek W, Wiegand T, Muller KR. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ITU J ICT Discov. 2017;Special Issue 1:39-48.

[4]. Davis M, Park H. Adoption of AI-augmented payment systems by consumers: Behavioral factors and implications for user experience. Rev Consum Technol. 2023;25(1):34-56.

[5]. Garcia R, Thompson A. A comparative examination of the regulatory frameworks governing artificial intelligence within the financial services industry. Financ Regul Q. 2024;31(2):112-34.

[6]. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54(6):1-35. Available from: https://doi.org/10.1145/3457607

[7]. Jensen MC, Meckling WH. Theory of the firm: Managerial behavior, agency costs and ownership structure. J Financ Econ. 1976;3(4):305-60. Available from: https://doi.org/10.1016/0304-405X(76)90026-X

[8]. Wilson J, Garcia M. Ethical considerations and methodological adjustments concerning access limitations in private technology research. Technol Res Ethics. 2023;5(4):167-84.

[9]. Anderson J, Kumar S. A multi-dimensional investigation of stakeholders perspectives on the application of artificial intelligence in financial services. Int J Bank Finance. 2023;41(2):123-45.

[10]. DiMaggio PJ, Powell WW. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. Am Sociol Rev. 1983;48(2):147-60. Available from: https://doi.org/10.2307/2095101

[11]. Thompson A, Lee D. Operational efficiency analysis in AI-enhanced payment systems: A cost-benefit assessment. Oper Res Finance. 2024;20(2):89-111.

[12]. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3(2):77-101.

Available from: https://doi.org/10.1191/1478088706qp063oa
[13]. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135-44. Available from: https://doi.org/10.1145/2939672.2939778
[14]. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.
[15]. Quantiphi. The AI revolution in payments: Improving speed, security, and consumer experience [Internet]. 2024. Available from: https://quantiphi.com/blog/the-ai-revolution-in-payments-enhancing-speed-security-and-customer-experience/
[16]. McKinsey and Company. Financial technology and the future of finance: The forthcoming upheaval [Internet]. McKinsey Global Institute; 2023. Available from: https://www.mckinsey.com/industries/financial-services/our-insights/fintech-and-the-future-of-finance
[17]. IBM. Artificial intelligence in fraud detection inside the banking sector [Internet]. IBM Technology; 2025. Available from: https://www.ibm.com/think/topics/ai-fraud-detection-in-banking
[18]. ResearchGate. Ethical and regulatory concerns associated with AI implementation in the financial services sector: A global perspective [Internet]. 2025. Available from: https://www.researchgate.net/publication/392062990
[19]. Johnson K, Martinez L, Thompson R. Regulatory policy development for artificial intelligence in financial services: Issues of implementation and enforcement. Financ Regul Rev. 2024;33(1):12-34.
[20]. Deloitte. The future of payments: The boundaries of tomorrow [Internet]. Deloitte Financial Services; 2023. Available from: https://www2.deloitte.com/us/en/pages/financial-services/articles/future-of-payments.html
[21]. Teece DJ. Explicating dynamic capabilities: The nature and microfoundations of sustainable enterprise performance. Strateg Manag J. 2007;28(13):1319-50. Available from: https://doi.org/10.1002/smj.640
[22]. Barney JB. Firm resources and sustained competitive advantage. J Manag. 1991;17(1):99-120. Available from: https://doi.org/10.1177/014920639101700108
[23]. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Manage Sci. 2000;46(2):186-204.
[24]. Intersog. AI in payment systems in 2025: The future of intelligent transactions [Internet]. 2025. Available from: https://intersog.com/blog/strategy/ai-in-payment-systems-in-2025-the-future-of-intelligent-transactions/
[25]. Number Analytics. Ten innovative digital payment applications improving customer experience [Internet]. 2025. Available from: https://www.numberanalytics.com/blog/innovative-digital-payment-customer-experience