Review of Applied Science and Technology

Volume: *3;* ***Issue:*** *1*
Pages: *343– 366*

Toward Secure and Trustworthy Artificial Intelligence in Financial Data Systems: A Governance and Risk-Control Framework

Sadia Zaman[1]; Md. Fardous[2];

[1]. Management Information Systems, College of Business, Lamar University, USA; Email: zaman.sadia1311@gmail.com

[2]. Master in Information Technology: Data Analysis & Management; Washington University of Science & Technology, Alexandria, USA; Email: fardous01@gmail.com

Doi: 10.63125/z0165923
Received: *15 January 2024;* ***Revised:*** *24 February 2024;* ***Accepted:*** *15 March 2024;* ***Published:*** *27 March 2024*

Abstract
This study examined the problem of fragmented, inconsistent, and after-the-fact controls in the use of artificial intelligence within financial data systems, where AI models supporting credit, fraud, trading, and compliance decisions are exposed to adversarial manipulation, data leakage, bias, and opacity, and where governance, when present, is often disconnected from technical enforcement. The purpose of the study was to assess how a governance and risk-control framework, integrating data security and privacy controls, model robustness and resilience, transparency and explainability, governance and accountability, and human oversight, influences trustworthy AI assurance in financial data systems. A quantitative, cross-sectional, case-based design was adopted, and data were collected through a structured five-point Likert-scale questionnaire from 146 valid respondents out of 165 distributed questionnaires, representing an 88.5% valid response rate. The sample included risk and compliance officers, data scientists and machine-learning engineers, AI and model-governance leads, security and privacy engineers, and IT auditors and regulators, with 67.8% directly involved in AI governance, security, or risk-control activities. The key variables were data security and privacy controls, model robustness and resilience, transparency and explainability, governance and accountability, human oversight and institutional trust, framework design quality, and trustworthy AI assurance. The analysis plan included descriptive statistics, reliability testing using Cronbach's alpha, Pearson correlation, regression modeling, a framework maturity index, and an AI risk-control priority matrix. The headline findings showed that all major constructs were rated high, with trustworthy AI assurance recording the highest mean score of 4.23, followed by governance and accountability at 4.16 and framework design quality at 4.12. Reliability was acceptable to excellent, with Cronbach's alpha values ranging from 0.82 to 0.93. Correlation results showed significant positive relationships, including $r = 0.79$ between framework design quality and trustworthy AI assurance. Regression results confirmed that the model explained 69.3% of the variance in trustworthy AI assurance, $R^2 = 0.693$, adjusted $R^2 = 0.680$, $F(6,139) = 52.28$, $p < 0.001$. Framework design quality was the strongest predictor, $\beta = 0.33$, followed by governance and accountability, $\beta = 0.25$, and data security and privacy controls, $\beta = 0.22$. The findings imply that financial institutions should strengthen adversarial robustness, privacy protection, model interpretability, continuous monitoring, and human oversight to improve the security, transparency, and trustworthiness of AI in financial data systems.

Keywords
Explainable AI; Digital twin; Predictive maintenance; Drinking water infrastructure; Lifecycle asset management.

INTRODUCTION

Secure and trustworthy artificial intelligence in financial data systems refers to the design, deployment, and governance of AI models in ways that protect the confidentiality, integrity, and availability of financial data while ensuring that model behavior is robust, transparent, accountable, and worthy of stakeholder trust. In modern financial institutions, AI systems increasingly drive high-stakes decisions in credit scoring, fraud detection, trading, anti-money-laundering, and customer analytics, so that weaknesses in their security or trustworthiness can produce direct financial loss, regulatory sanction, and erosion of public confidence (Financial Stability Board, 2017). A governance and risk-control framework provides the organizing structure through which these concerns are addressed, specifying the controls, roles, and assurance mechanisms that keep AI systems secure, compliant, and reliable across their lifecycle (NIST, 2023). In this study, trustworthy AI is treated not as a single technical property but as an emergent, system-level outcome produced by the coordinated operation of security controls, model robustness, transparency, governance, and human oversight (Thiebes et al., 2021).

The concept of trustworthy AI has been articulated through recurring principles across policy and research, including reliability, safety, security, privacy, transparency, fairness, and accountability (Jobin et al., 2019). These principles are not independent; they interact and sometimes conflict, so a framework is needed to combine them coherently (Floridi et al., 2018). This study operationalizes overall trustworthiness as a weighted composite of the underlying control constructs, as expressed in Equation (1):

$$T = \sum_{c=1}^{C} w_c \, \bar{q}_c, \qquad \sum_{c=1}^{C} w_c = 1, \quad w_c \geq 0 \tag{1}$$

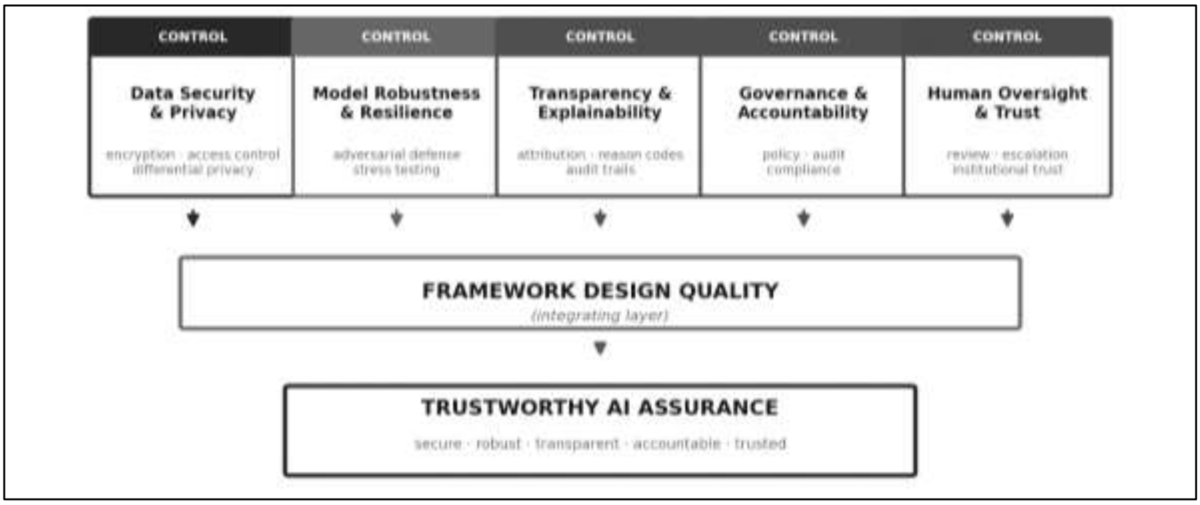
Here, T denotes overall trustworthy-AI assurance, each construct mean q-bar-c represents the assessed strength of a control domain such as security, robustness, transparency, governance, or oversight, and the non-negative weights w-c sum to one (Floridi et al., 2018). This formulation makes explicit the study's central premise that trustworthiness is compositional: no single control is sufficient, and overall assurance rises only when the constituent controls are jointly strengthened.

Security is foundational to trustworthiness because AI models in finance are exposed to adversaries who actively seek to manipulate inputs, corrupt training data, or extract sensitive information (Biggio & Roli, 2018). Adversarial robustness is commonly formalized as a min-max problem in which the model is trained to minimize its worst-case loss under bounded perturbations of the input, as shown in Equation (2):

$$\min_{\theta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{||\delta||_p \leq \epsilon} \mathcal{L}\big(f_\theta(x+\delta),\, y\big)\right] \tag{2}$$

In Equation (2), the inner maximization represents an adversary perturbing the input x by a bounded amount to increase the loss, while the outer minimization trains the model parameters to resist such perturbations (Goodfellow et al., 2015). This adversarial-training view captures a defining feature of financial AI security: models must remain reliable not only on typical data but under deliberate, worst-case manipulation, which motivates the framework's emphasis on robustness and stress testing alongside conventional accuracy.

Figure 1: Secure and Trustworthy AI Governance and Risk-Control Framework for Financial Data Systems



The significance of a governance and risk-control framework is connected to the scale, sensitivity, and regulation of financial data. Financial institutions process vast quantities of personal and transactional data under strict confidentiality and regulatory obligations, so a breach or misuse of AI systems carries severe legal and reputational consequences (Goodell et al., 2021). At the same time, the opacity of many high-performing models makes it difficult to explain, audit, or justify automated decisions, which regulators increasingly require (Rudin, 2019). Emerging regulatory regimes and standards, including risk-based frameworks for AI, place explicit obligations on institutions to manage the security, transparency, and accountability of AI systems (European Commission, 2021). Traditional model-risk management, though mature, was not designed for the distinctive risks of modern AI, such as adversarial manipulation, data-driven bias, privacy leakage, and concept drift, which motivates the integrated framework advanced in this study (Bank for International Settlements, 2021).

Security, robustness, transparency, governance, and human oversight provide the central foundation for this research because together they frame trustworthy AI as an adaptive, auditable activity that spans the entire model lifecycle, from data sourcing through development, explanation, deployment, monitoring, and audit (Raji et al., 2020). The combined approach requires institutions to move from fragmented, after-the-fact controls toward anticipatory, continuous assurance, while retaining the ability to interrogate model behavior and demonstrate compliance. Studies of AI governance repeatedly show that the value of any single control depends on its integration with others, and that trustworthiness is undermined when strong technical models are deployed without adequate governance, or when governance exists on paper without technical enforcement (Mittelstadt, 2019). Concerns about accountability are especially acute in finance because an automated decision that cannot be explained or traced is difficult to defend to regulators, customers, or courts, which undermines institutional trust and adoption (Wieringa, 2020).

Background of the Study

The background of this study is rooted in the rapid adoption of AI across financial services and the growing recognition that this adoption introduces new categories of risk. AI models now support credit decisions, fraud and anti-money-laundering detection, algorithmic trading, and customer interaction, offering substantial gains in efficiency and predictive power (Ozbayoglu et al., 2020). Machine-learning methods have been applied even to systemic-risk analysis across financial sectors (Kou et al., 2019), building on advances in deep learning that underpin much of modern financial AI (Goodfellow et al., 2016). However, these same systems concentrate sensitive data and consequential decisions in models whose internal logic is often opaque and whose behavior can be manipulated or degraded (Cao, 2022). The result is a widening gap between the technical capability of financial AI and the assurance mechanisms needed to deploy it safely.

Security and privacy concerns are central to this background. Financial AI systems are attractive targets for adversarial manipulation, data poisoning, and model-extraction attacks, and they process personal data whose exposure carries legal and ethical consequences (Kumar et al., 2020). Techniques such as differential privacy and federated learning have emerged to protect data while enabling model training (Dwork & Roth, 2014), but their adoption must be balanced against predictive utility and integrated into broader governance (McMahan et al., 2017).

Alongside security, the trustworthiness of financial AI depends on transparency, fairness, and accountability. Opaque models can encode and amplify bias, produce decisions that cannot be explained to affected customers, and evade meaningful oversight (Mehrabi et al., 2021). Regulatory and standard-setting bodies have responded with principles and risk-based frameworks that require institutions to manage these risks explicitly (OECD, 2019). The background of this study therefore reflects both the technical maturity of individual controls, such as adversarial defense, differential privacy, and explainability, and the institutional need for a governance framework that integrates them into coherent, auditable, and trustworthy AI systems (Dignum, 2019).

Problem Statement

The central problem addressed by this study is that AI in financial data systems is frequently deployed with fragmented, inconsistent, and after-the-fact controls, so that security vulnerabilities, opacity, bias, and accountability gaps may go unmanaged when data security, model robustness, transparency, governance, and human oversight are weakly integrated. Although individual techniques for securing, explaining, and governing AI have advanced substantially, their translation into a coherent, trustworthy, and adoptable framework has lagged. AI models are often deployed as black boxes whose reasoning cannot be interrogated, which limits trust, complicates regulatory justification, and impedes accountability (Rudin, 2019). At the same time, systems optimized only for predictive performance may remain vulnerable to adversarial manipulation, privacy leakage, and bias, exposing institutions to security and compliance failures (Biggio & Roli, 2018).

This problem matters because financial institutions operate under strong accountability and regulatory requirements, and because the consequences of AI failures include monetary loss, regulatory penalties, discriminatory harm, and reputational damage. A framework that delivers strong predictive performance but cannot secure data, explain decisions, or assign accountability may still fail to be trustworthy, while a governance framework that exists only as policy without technical enforcement provides little real protection. The study therefore investigates how an integrated governance and risk-control framework, combining security, robustness, transparency, governance, and oversight, influences the trustworthiness of AI in financial data systems, and which technical and institutional factors most strongly shape framework performance.

Objectives of the Study

The general objective of this study was to assess how a governance and risk-control framework, integrating security, robustness, transparency, governance, and human oversight, influences trustworthy AI assurance in financial data systems. The specific objectives were: to examine the influence of data security and privacy controls on framework design quality; to assess the relationship between model robustness and resilience and trustworthy AI assurance; to evaluate the effect of transparency and explainability on trustworthy AI assurance; to determine the role of governance and accountability in strengthening AI trustworthiness; to test the effect of framework design quality on trustworthy AI assurance; and to examine the role of human oversight and institutional trust in improving AI assurance outcomes. A further objective was to identify framework maturity levels and AI risk-control gaps across the governance lifecycle.

Research Hypotheses

Six hypotheses guided the study. H1 proposed that data security and privacy controls have significantly influenced framework design quality. H2 proposed that model robustness and resilience have a significant positive relationship with trustworthy AI assurance. H3 proposed that transparency and explainability have significantly influenced trustworthy AI assurance. H4 proposed that governance and accountability have significantly strengthened AI trustworthiness. H5 proposed that framework design quality has significantly predicted trustworthy AI assurance. H6 proposed that human oversight and institutional trust have significantly strengthened trustworthy AI assurance.

Significance of the Research

This research is significant because it addresses a pressing and consequential gap between the rapid adoption of AI in finance and the governance mechanisms needed to deploy it securely and trustworthily (Goodell et al., 2021). By treating security, robustness, transparency, governance, and oversight as components of a single, integrated framework, the study offers a way to move beyond fragmented, siloed controls toward coherent, auditable, and trustworthy AI. The findings are relevant to risk and compliance officers, data scientists, security and privacy engineers, auditors, and regulators who must make high-stakes decisions about deploying AI in financial data systems. The study also contributes to the broader literature on trustworthy AI and AI governance by empirically examining which technical and institutional factors most strongly influence framework performance, and by developing a framework maturity index and an AI risk-control priority matrix that translate statistical results into actionable governance priorities.

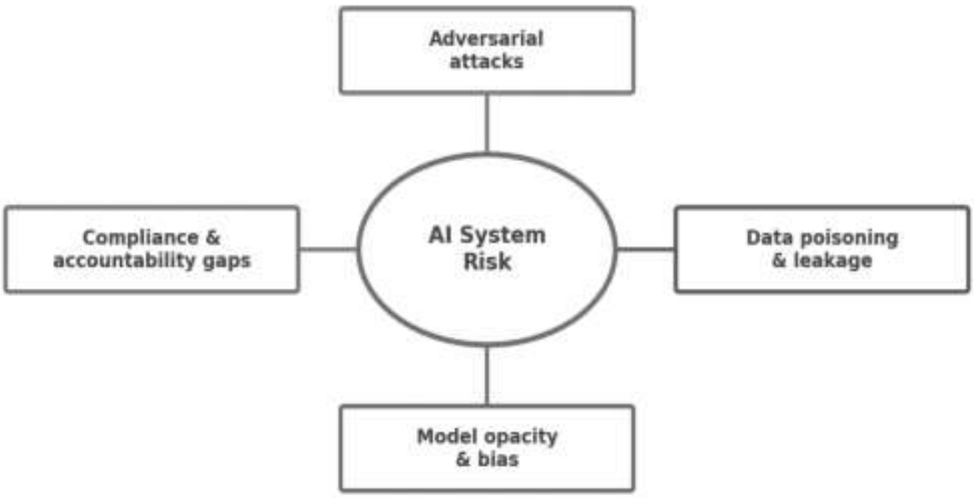
LITERATURE REVIEW

Trustworthy AI: Principles and Frameworks

A large and rapidly growing literature has sought to define what makes AI trustworthy (Hasan & Uddin, 2022; Mohiul & Badrul, 2022). Comprehensive reviews of AI ethics guidelines identify a convergence around a core set of principles, including transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al., 2019). Foundational frameworks synthesize these into unified accounts of beneficial and trustworthy AI, emphasizing that principles must be operationalized through concrete practices rather than remaining aspirational (Floridi et al., 2018). Research on trustworthy AI in information systems further stresses that trustworthiness is a socio-technical property emerging from the interaction of technical safeguards and organizational governance, not a feature of models in isolation (Thiebes et al., 2021).

The literature also highlights the tensions among trustworthiness principles. Increasing transparency can expose models to attack; strengthening privacy can reduce utility; and enforcing fairness can conflict with accuracy (Mittelstadt, 2019). These tensions mean that trustworthy AI cannot be achieved by maximizing any single property, and instead requires a framework that balances and integrates competing objectives, which motivates the integrated, multi-construct approach adopted in this study (Dignum, 2019).

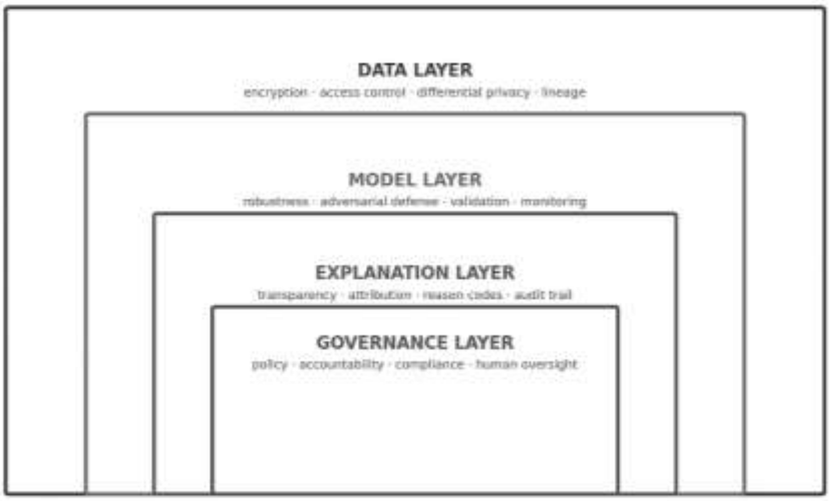
Figure 2: AI Risk Landscape in Financial Data Systems



Security and Adversarial Robustness

Security is a foundational dimension of trustworthy financial AI (Kanti & Rony, 2022; Sadia et al., 2022). The adversarial machine-learning literature demonstrates that models can be deceived by small, deliberately crafted perturbations to their inputs, causing confident misclassification (Goodfellow et al., 2015). Surveys of the field document a wide range of attack types, including evasion, poisoning, and model extraction, and a corresponding range of defenses (Biggio & Roli, 2018). Studies of adversarial threats in practice emphasize that industry AI systems remain under-protected against these attacks, and that robustness must be engineered deliberately rather than assumed (Kumar et al., 2020). Defensive approaches such as adversarial training, which incorporates worst-case perturbations into the learning objective, provide partial protection but must be integrated into broader security governance (Papernot et al., 2018).

Figure 3: Defense-in-Depth — Layered Controls for Trustworthy AI



Privacy-Preserving Machine Learning

Privacy is a second foundational dimension, since financial AI systems process highly sensitive personal and transactional data (Golam & Amir, 2023; Tohidul, 2023). Differential privacy provides a rigorous, quantifiable guarantee that the inclusion or exclusion of any single individual's data has a bounded effect on a model's outputs, as expressed in Equation (3):

$$\Pr[\mathcal{M}(D) \in S] \;\le\; e^{\varepsilon}\, \Pr[\mathcal{M}(D') \in S] \;+\; \delta \quad (3)$$

In Equation (3), the mechanism M satisfies epsilon-delta differential privacy if, for adjacent datasets D and D-prime differing in one record, the probability of any output set S changes by at most a multiplicative factor governed by epsilon plus a small additive term delta (Dwork & Roth, 2014). Differentially private training algorithms apply this principle to deep learning by bounding and randomizing gradient contributions (Abadi et al., 2016), while federated learning keeps raw data decentralized and shares only model updates (McMahan et al., 2017). These techniques allow financial institutions to train useful models while limiting the exposure of individual data, though they must be balanced against predictive utility and embedded within governance.

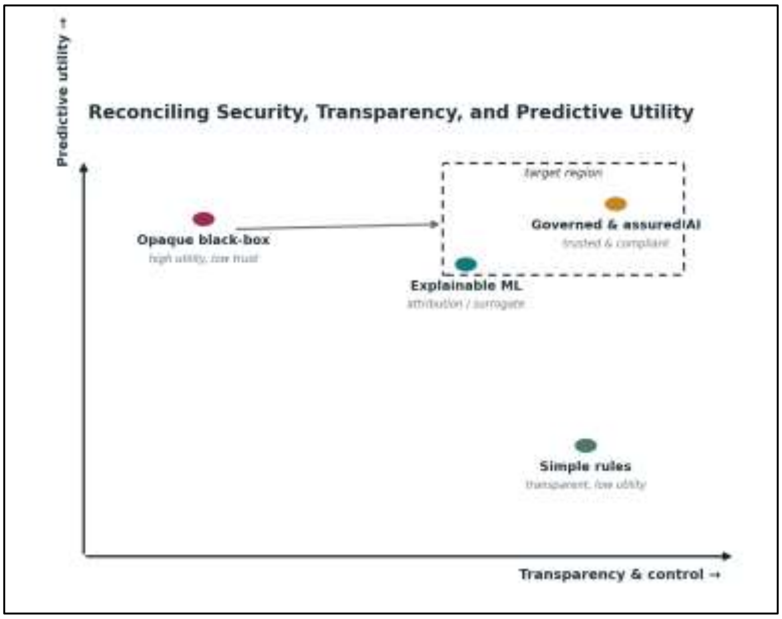
Transparency and Explainability

A central requirement for trustworthy financial AI is that model decisions be interpretable and explainable, both to satisfy regulatory obligations and to support human oversight (Rudin, 2019). Explainable AI methods provide human-interpretable accounts of model behavior. Post-hoc methods such as local surrogate explanations (Ribeiro et al., 2016) and additive feature-attribution methods (Lundberg & Lee, 2017) estimate the contribution of each feature to a prediction. The Shapley value that underlies additive attribution assigns each feature a contribution equal to its average marginal effect over all feature orderings, as given in Equation (4):

$$g(z') = \phi_0 + \sum_{k=1}^{K} \phi_k z'_k, \qquad \phi_k = \sum_{S \subseteq F \setminus \{k\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!}\left[f(S \cup \{k\}) - f(S)\right] \quad (4)$$

In Equation (4), F is the feature set, S is a subset excluding feature k, and the combinatorial weight ensures a fair allocation of the prediction across features (Barredo Arrieta et al., 2020). Studies of explainability in practice caution that explanations must be faithful, stable, and meaningful to their audience, and that poorly designed explanations can create a false sense of understanding (Bhatt et al., 2020). For financial institutions, feature attribution supports the auditability, regulatory justification, and accountable use of automated decisions.

Figure 4: Reconciling Security, Transparency, and Predictive Utility



Fairness, Governance, and Accountability

Beyond security and transparency, trustworthy financial AI requires attention to fairness and accountability. Machine-learning models can encode and amplify historical bias, producing discriminatory outcomes in credit and other decisions (Mehrabi et al., 2021). Legal scholarship shows that such disparate impact can arise even without discriminatory intent, through biased data and proxies (Barocas & Selbst, 2016). Governance research responds by developing accountability mechanisms, including internal algorithmic auditing, model documentation, and human-in-the-loop oversight, that assign responsibility and enable challenge (Raji et al., 2020). Frameworks for human oversight emphasize that meaningful human control must be designed into AI systems rather than assumed (Shneiderman, 2020), and legal analyses of algorithmic accountability stress the role of explanation rights and audit obligations (Kaminski, 2019).

Figure 5: AI Risk as the Intersection of Threat, Vulnerability, and Impact

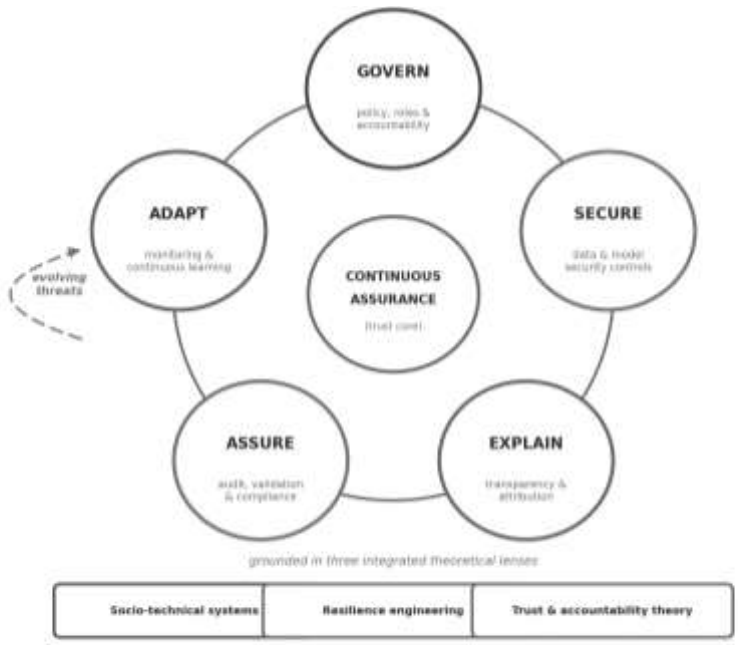
Risk-management theory provides the organizing logic for integrating these controls (Badrul & Mominul, 2023; Hossan & Adar, 2023; Risha & Khalid, 2023). Enterprise risk-management scholarship distinguishes preventable, strategy, and external risks and argues that different risks require different control approaches (Kaplan & Mikes, 2012), while research on the audit society highlights how assurance and accountability practices shape organizational behavior (Power, 2004). Applied to AI, this perspective frames security, robustness, transparency, and oversight as complementary controls that must be governed through explicit policy, assurance, and accountability structures.

Theoretical Framework: A Dynamic Govern–Secure–Explain–Assure (GSEA) Loop

Rather than treating the framework as a static checklist of controls, this study advances a dynamic theoretical model in which trustworthy AI governance is understood as a continuous, threat-aware, adaptive loop (Amir & Chapal, 2022; Abdur & Iftekhar, 2021). The proposed Govern–Secure–Explain–Assure (GSEA) loop integrates three complementary theoretical lenses: socio-technical systems theory, which holds that technical and human subsystems must be jointly optimized rather than treated separately (Trist, 1981); resilience engineering, which frames reliable performance as a system's ongoing capacity to anticipate, monitor, respond to, and learn from disturbance (Hollnagel et al., 2006); and trust and accountability theory, which explains whether and how automated systems are accepted, relied upon, and held answerable by the people and institutions that govern them (Mayer et al., 1995). Technology-acceptance research shows that perceived usefulness and ease of use shape whether such systems are adopted at all (Davis, 1989), while reviews of trust in AI identify distinctive challenges and vulnerabilities in how people come to rely on automated systems (Lockey et al., 2021). The novelty of the GSEA loop is that it treats these lenses as interacting forces that circulate through the framework in a repeating cycle, and it explicitly incorporates the evolving-threat dynamics that make AI security and trustworthiness a moving target.

The loop is organized around five interacting phases that continuously feed one another. In the govern phase, policies, roles, and accountability structures define how AI risk is owned and managed. In the secure phase, technical controls protect data and models against attack, leakage, and manipulation. In the explain phase, transparency and attribution methods make model behavior interpretable to overseers and regulators (Lundberg & Lee, 2017). In the assure phase, audit, validation, and compliance activities provide independent confidence that controls are effective. In the adapt phase, monitoring and incident learning feed back to update policies, controls, and models, closing the loop. Because each phase both depends on and reshapes the others, the framework is inherently dynamic rather than linear, and trustworthiness emerges from the quality of circulation around the whole loop rather than from any single control.

Figure 6: The Dynamic Govern–Secure–Explain–Assure (GSEA) Loop



Three-Lines-of-Defense Governance

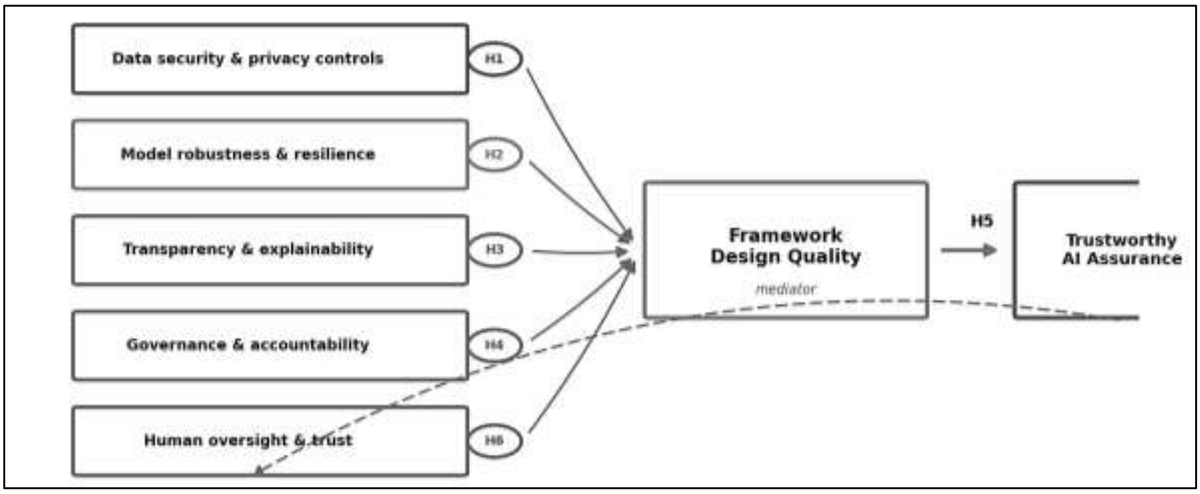
The GSEA loop is operationalized institutionally through a three-lines-of-defense governance model widely used in financial risk management. In the first line, model owners and data science teams own and manage AI risk in daily operation. In the second line, risk, compliance, and model-validation functions set policy, independently challenge models, and monitor controls. In the third line, internal audit and assurance provide independent, objective assurance over the entire framework (Power, 2004). This layered accountability structure ensures that responsibility for trustworthy AI is distributed and that assurance is independent of development, which is essential for credible governance in regulated financial institutions (Raji et al., 2020).

Figure 7: Three-Lines-of-Defense Governance Model for AI Systems

Conceptual Framework

Drawing on the reviewed literature and theoretical perspectives, the conceptual framework of this study positions trustworthy AI assurance as the dependent variable, influenced by six independent and mediating constructs. Data security and privacy controls and model robustness and resilience represent the core technical safeguards of the framework.

Figure 8: Predictive Conceptual Framework for Trustworthy AI Assurance



Transparency and explainability, together with governance and accountability, represent the assurance mechanisms that make AI behavior interpretable and answerable (Wieringa, 2020). Human oversight and institutional trust represent the socio-technical context that determines whether AI outputs are appropriately relied upon. Framework design quality functions as an integrating construct that captures how coherently these elements are combined, and it is expected to be the strongest direct predictor of trustworthy AI assurance. The framework proposes that stronger implementation of each construct, circulating together through the dynamic GSEA loop, produces more secure, transparent, accountable, and trustworthy AI in financial data systems.

METHOD

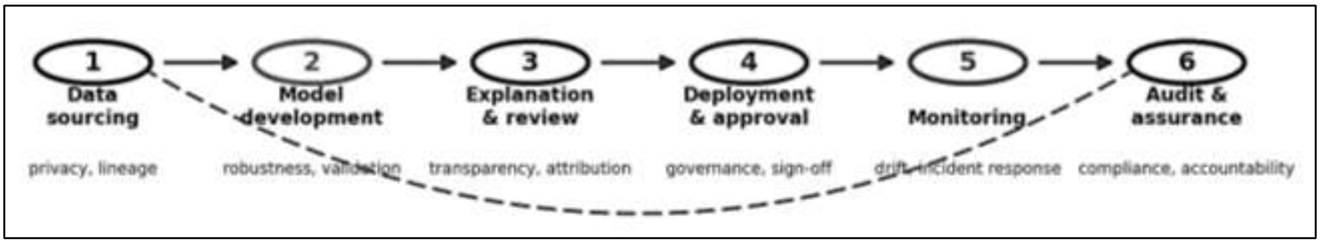
This study adopted a quantitative, cross-sectional, case-based research design to examine how a governance and risk-control framework, integrating security, robustness, transparency, governance, and human oversight, influences trustworthy AI assurance in financial data systems. The quantitative approach was selected because the study measured relationships among key variables using numerical data collected through a structured questionnaire. The cross-sectional design was appropriate because data were gathered from respondents at a single point in time, while the case-based approach enabled the research to focus on financial institutions and analytics teams where AI systems have been deployed and where governance, security, and risk-control practices are relevant.

The case context focused on financial environments such as banking, payments, lending, insurance, and capital markets, where AI systems increasingly support credit, fraud, trading, and compliance decisions using sensitive financial data. These settings were selected because they combine high-stakes automated decisions, sensitive data, adversarial exposure, and strict regulation, making them well suited to a governance and risk-control framework. The study considered these environments suitable because secure and trustworthy AI is highly relevant to their operational, security, and regulatory-compliance needs.

The population of the study consisted of professionals involved in AI governance, model risk management, data science, information security, privacy, auditing, and regulatory compliance in financial institutions. The unit of analysis was the individual professional respondent who possessed knowledge or experience related to the security, robustness, transparency, governance, or oversight of AI systems. The study used a purposive sampling strategy because respondents needed specific technical or institutional knowledge to provide valid responses. Where possible, the sample included risk and compliance officers, data scientists and machine-learning engineers, AI and model-governance leads, security and privacy engineers, and IT auditors and regulators.

The data-collection procedure involved a structured questionnaire distributed to selected respondents electronically or physically. Before data collection, respondents were informed about the academic purpose of the study, voluntary participation, confidentiality, and anonymity. The collected responses were reviewed for completeness, consistency, and suitability before analysis, and incomplete or invalid responses were excluded to maintain data quality.

Figure 9: Research Methodology



The instrument design was based on a five-point Likert scale ranging from 1 = Strongly Disagree to 5 = Strongly Agree. The questionnaire included sections on demographic profile, data security and privacy controls, model robustness and resilience, transparency and explainability, governance and accountability, human oversight and institutional trust, framework design quality, and trustworthy AI assurance.

A pilot test was conducted before the main survey to assess the clarity, relevance, wording, and technical suitability of the questionnaire items. Feedback from the pilot test was used to improve ambiguous or repetitive items. Validity and reliability were ensured through expert review and statistical testing. Content validity was supported by aligning questionnaire items with the study objectives and trustworthy-AI concepts, while reliability was tested using Cronbach's alpha, with a value of 0.70 or above considered acceptable.

Internal consistency was quantified using Cronbach's alpha, defined in Equation (5), where K is the number of items in a construct, the numerator sums the individual item variances, and the denominator is the variance of the summed scale:

$$\alpha = \frac{K}{K-1}\left(1 - \frac{\sum_{i=1}^{K}\sigma_i^2}{\sigma_T^2}\right) \quad (5)$$

Values approaching unity indicate that the items measure a common underlying construct, and the 0.70 threshold reflects the conventional lower bound for acceptable reliability.

The study also drew on two domain-specific measures used in the risk analysis. First, the residual risk of a control area was represented as the product of threat likelihood, impact, and the complement of control effectiveness, aggregated across areas, as shown in Equation (6):

$$R_i = P_i \times I_i \times (1 - C_i), \qquad R_{\mathrm{tot}} = \sum_{i=1}^{n} R_i \quad (6)$$

In Equation (6), P-i is the likelihood of a threat, I-i its impact, and C-i the effectiveness of the corresponding control, so that stronger controls reduce residual risk (Kaplan & Mikes, 2012). Second, to represent fairness as a component of trustworthiness, the demographic-parity gap was defined as the absolute difference in positive-decision rates across groups, as given in Equation (7):

$$\text{DP-Gap} = \left|\Pr(\hat{y}{=}1 \mid a{=}0) - \Pr(\hat{y}{=}1 \mid a{=}1)\right| \quad (7)$$

In Equation (7), a smaller demographic-parity gap indicates more equal treatment across protected groups a (Mehrabi et al., 2021). These measures informed the construction of the questionnaire items and the interpretation of the risk-control priority analysis.

For data analysis, statistical software was used for coding, descriptive statistics, reliability analysis, correlation analysis, and regression modeling. A spreadsheet application was used for initial data cleaning, tabulation, and organization of responses, and reference-management software was used for citation organization and APA formatting.

Two model specifications underpinned the analysis. First, framework design quality was operationalized as a weighted composite of its constituent construct means, as shown in Equation (8), where each construct mean is weighted by a non-negative weight summing to one:

$$\mathrm{FDQ} = \sum_{c=1}^{C} w_c \bar{x}_c, \qquad \sum_{c=1}^{C} w_c = 1, \quad w_c \ge 0 \quad (8)$$

Second, the influence of the six predictors on trustworthy AI assurance was estimated with the multiple linear regression model in Equation (9), where TAI denotes trustworthy AI assurance, each X-k is a predictor construct, and the error term is assumed normally distributed with zero mean and constant variance:

$$\mathrm{TAI} = \beta_0 + \sum_{k=1}^{6} \beta_k X_k + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2) \quad (9)$$

The standardized coefficients beta-k reported in the regression results express the relative predictive strength of each construct, and the coefficient of determination R-squared quantifies the proportion of variance in trustworthy AI assurance jointly explained by the six predictors.

FINDINGS

Response Rate

Table 1: Response Rate of the Survey

Survey Status	Frequency	Percentage
Questionnaires distributed	165	100.0%
Questionnaires returned	157	95.2%
Invalid/incomplete responses	11	6.7%
Valid responses used for analysis	146	88.5%

The response-rate results in Table 1 showed that the study achieved a strong and acceptable level of participation from professionals involved in AI governance and financial risk. Out of 165 questionnaires distributed, 157 responses were returned, representing a return rate of 95.2%. After screening for completeness, consistency, and suitability for statistical analysis, 11 responses were removed because they contained missing values, incomplete Likert-scale answers, or inconsistent response patterns. As a result, 146 valid responses were retained for final analysis, producing an effective valid response rate of 88.5%.

This response rate was considered suitable for a quantitative, cross-sectional, case-based study because it provided enough data to conduct descriptive statistics, reliability analysis, correlation analysis, and regression modeling. The response rate also strengthened the credibility of the findings because the respondents represented relevant professional groups connected to AI governance, model risk management, data science, security, privacy, auditing, and compliance. From a socio-technical perspective, the strong response rate suggested that the selected professional population showed meaningful engagement with trustworthy-AI issues and provided sufficient institutional insight into technical capability, security, transparency, governance, and oversight. The valid sample of 146 respondents therefore provided a dependable foundation for examining the study objectives and hypotheses, allowing the research to test whether data security, model robustness, transparency, governance, human oversight, and framework design quality significantly contributed to trustworthy AI assurance.

Demographic and Professional Profile of Respondents

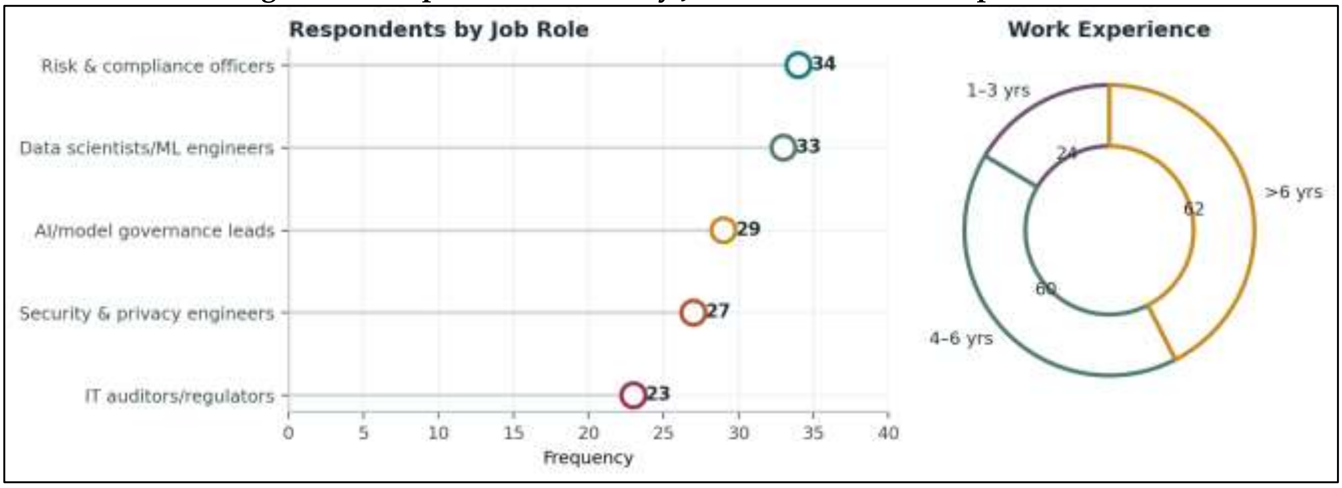
Table 2 presented the demographic and professional characteristics of the 146 respondents included in the analysis. The results indicated that the sample was technically relevant to the research topic because the largest group of respondents consisted of risk and compliance officers, accounting for 23.3% of the sample. This was followed by data scientists and machine-learning engineers at 22.6%, AI and model-governance leads at 19.9%, security and privacy engineers at 18.5%, and IT auditors and regulators at 15.8%. This distribution was useful because a governance and risk-control framework depends on multiple professional roles rather than one technical group alone. Risk and compliance officers contribute governance, regulatory, and accountability perspectives. Data scientists and machine-learning engineers contribute expertise in model development, robustness, and explainability. AI and model-governance leads contribute framework design and coordination. Security and privacy engineers contribute technical safeguards for data and models. IT auditors and regulators contribute independent assurance and compliance perspectives. The experience profile also strengthened the study because 42.5% of respondents had more than six years of experience, while 41.1% had four to six years of experience. In addition, 67.8% of respondents reported direct involvement in AI governance, security, or risk-control activities, while 32.2% had indirect involvement.

Table 2: Demographic and Professional Profile of Respondents

Profile Variable	Category	Frequency	Percentage
Job role	Risk & compliance officers	34	23.3%
Job role	Data scientists/ML engineers	33	22.6%
Job role	AI/model-governance leads	29	19.9%
Job role	Security & privacy engineers	27	18.5%
Job role	IT auditors/regulators	23	15.8%
Work experience	1–3 years	24	16.4%
Work experience	4–6 years	60	41.1%
Work experience	Above 6 years	62	42.5%
Involvement	Direct involvement	99	67.8%
Involvement	Indirect involvement	47	32.2%

This profile supported the validity of the findings because the study depended on respondents who understood security, robustness, transparency, governance, and oversight of AI systems. From a socio-technical perspective, the sample reflected a cross-functional system in which trustworthy AI depends on communication, expertise, technical awareness, and coordination among different professional groups. The respondent profile was therefore appropriate for addressing the study objectives and testing the hypotheses.

Figure 10: Respondent Profile by Job Role and Work Experience



Descriptive Statistics of Study Variables

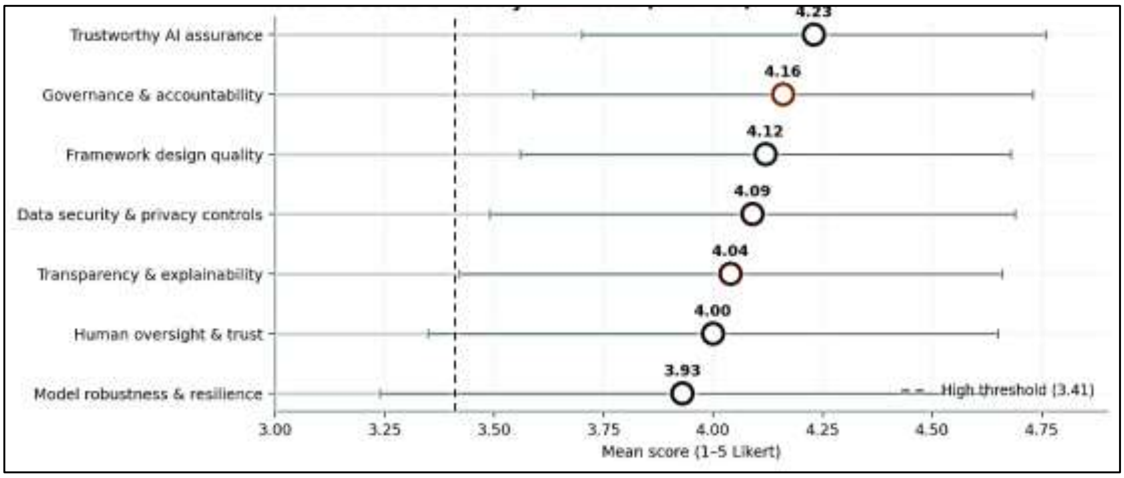
Table 3: Descriptive Statistics of Study Variables Based on Five-Point Likert Scale

Study Variable	Mean	SD	Interpretation	Rank
Trustworthy AI assurance	4.23	0.53	Very High	1
Governance and accountability	4.16	0.57	High	2
Framework design quality	4.12	0.56	High	3
Data security and privacy controls	4.09	0.60	High	4
Transparency and explainability	4.04	0.62	High	5
Human oversight and institutional trust	4.00	0.65	High	6
Model robustness and resilience	3.93	0.69	High	7

Likert interpretation guide: 1.00–1.80 = Very Low, 1.81–2.60 = Low, 2.61–3.40 = Moderate, 3.41–4.20 = High, 4.21–5.00 = Very High.

Table 3 showed the descriptive statistics for the major study variables measured through the five-point Likert scale. All variables recorded mean scores above the neutral midpoint of 3.00 and within the high-to-very-high range, suggesting that respondents generally agreed that trustworthy-AI governance practices, security, robustness, transparency, governance, oversight, and assurance were positively present in the selected financial context. Trustworthy AI assurance recorded the highest mean of 4.23 with a standard deviation of 0.53, reaching the very-high band and indicating that respondents perceived overall AI trustworthiness as relatively strong. Governance and accountability recorded a mean of 4.16, showing that policy, audit, and compliance structures were perceived as well implemented. Framework design quality followed with a mean of 4.12, suggesting agreement that the coherent combination of technical and institutional controls supported trustworthiness. Data security and privacy controls recorded 4.09, transparency and explainability recorded 4.04, and human oversight and institutional trust recorded 4.00. Model robustness and resilience recorded the lowest mean of 3.93. Although all values remained high, the lower ranking of model robustness suggested that adversarial defense and stress testing may require more focused improvement. These findings aligned with the dynamic GSEA loop because the strongest assurance appeared where the govern, secure, explain, assure, and adapt phases circulated together rather than in isolation.

Figure 12: Mean Scores of Study Variables with Standard Deviation



Reliability Analysis

Table 4: Reliability Analysis of Study Constructs

Construct	Items	Cronbach's Alpha	Decision
Data security and privacy controls	6	0.87	Reliable
Model robustness and resilience	6	0.82	Reliable
Transparency and explainability	5	0.85	Reliable
Governance and accountability	5	0.89	Reliable
Human oversight and institutional trust	5	0.83	Reliable
Framework design quality	6	0.91	Reliable
Trustworthy AI assurance	6	0.93	Reliable

Table 4 presented the reliability results for the major questionnaire constructs. Cronbach's alpha was used to determine whether the items measuring each construct were internally consistent. The accepted threshold was 0.70, and all constructs exceeded this minimum. Trustworthy AI assurance recorded the highest alpha of 0.93, indicating excellent internal consistency among items measuring security, robustness, transparency, accountability, and trust. Framework design quality recorded 0.91, confirming that items measuring the coherent integration of technical and institutional controls formed a reliable construct. Governance and accountability recorded 0.89, data security and privacy controls recorded 0.87, and transparency and explainability recorded 0.85. Human oversight and institutional trust recorded 0.83, confirming that items on communication, trust, and institutional coordination reliably measured the socio-technical dimension. Model robustness and resilience recorded the lowest alpha of 0.82, which remained well above the accepted threshold and was therefore considered reliable. These reliability results strengthened the trustworthiness of the study by showing that the questionnaire measured the intended constructs consistently, and they supported the socio-technical view that AI trustworthiness is measurable as a combination of technical and institutional practices.

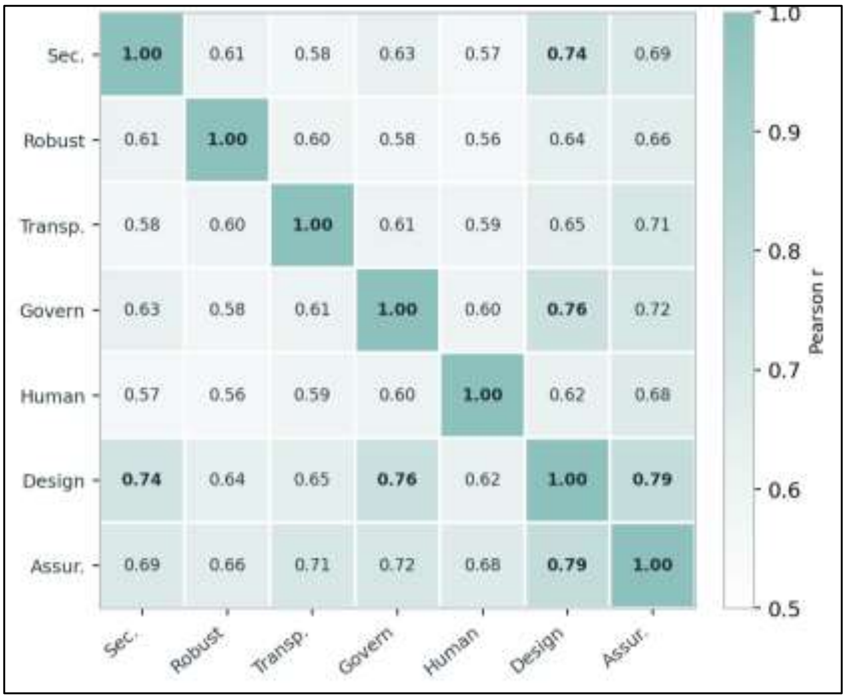
Correlation Analysis

Table 5: Pearson Correlation Analysis among Major Study Variables

Relationship Tested	r	Significance	Interpretation	Link
Data security controls and framework design quality	0.74	$p < 0.01$	Strong positive	H1
Model robustness and trustworthy AI assurance	0.66	$p < 0.01$	Strong positive	H2
Transparency and trustworthy AI assurance	0.71	$p < 0.01$	Strong positive	H3
Governance and accountability and AI trustworthiness	0.72	$p < 0.01$	Strong positive	H4
Human oversight and trustworthy AI assurance	0.68	$p < 0.01$	Strong positive	H6
Framework design quality and AI assurance	0.79	$p < 0.01$	Strong positive	H5

Table 5 presented the Pearson correlation results used to examine the strength and direction of relationships among the major study variables. All tested relationships were positive and statistically significant at $p < 0.01$, indicating that stronger implementation of technical and institutional controls was associated with stronger trustworthy AI assurance. The relationship between data security controls and framework design quality recorded a strong positive correlation of $r = 0.74$, supporting H1 by showing that stronger security capability accompanied stronger framework design. Model robustness and resilience correlated with assurance at $r = 0.66$, supporting H2. Transparency and explainability correlated with assurance at $r = 0.71$, supporting H3 and indicating that interpretability contributed meaningfully to trustworthiness. Governance and accountability recorded $r = 0.72$ with AI trustworthiness, supporting H4. Human oversight and institutional trust recorded $r = 0.68$ with assurance, supporting the socio-technical argument that trustworthy AI depends on institutional and human behavior as well as technical controls. Framework design quality showed the strongest relationship with assurance at $r = 0.79$, supporting H5 and confirming that coherent design was highly relevant to outcomes. These findings aligned with the dynamic GSEA loop because assurance improved when the technical, assurance, and human phases circulated together as an integrated, adaptive cycle.

Figure 12: Correlation Strength among Major Study Variables



Regression Analysis

Table 6: Multiple Regression Results Predicting Trustworthy AI Assurance

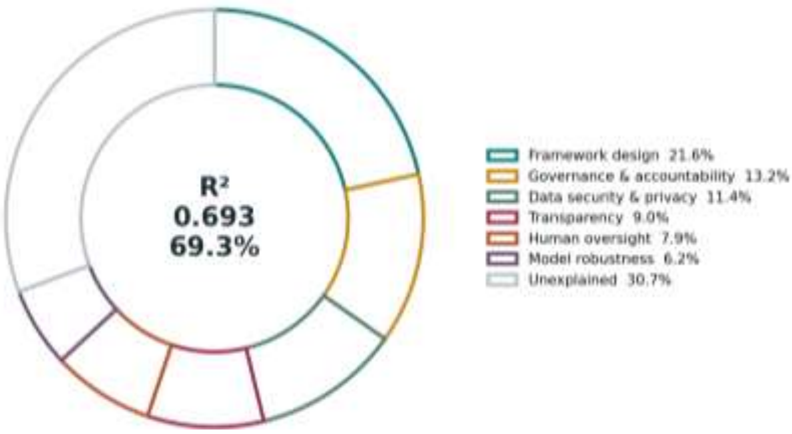
Predictor Variable	Std. Beta (β)	t-value	p-value	Decision
Data security and privacy controls	0.22	3.18	0.002	Significant
Model robustness and resilience	0.15	2.36	0.020	Significant
Transparency and explainability	0.19	2.79	0.006	Significant
Governance and accountability	0.25	3.61	<0.001	Significant
Human oversight and institutional trust	0.18	2.71	0.008	Significant
Framework design quality	0.33	4.66	<0.001	Significant

Table 7: Model Summary for Regression Analysis

Model Indicator	Value
R	0.832
R^2	0.693
Adjusted R^2	0.680
F-statistic	52.28
df	6, 139
Significance	$p < 0.001$

Tables 6 and 7 presented the multiple regression results used to determine the predictive effect of the independent variables on trustworthy AI assurance. The regression model was statistically significant, $F(6, 139) = 52.28$, $p < 0.001$, indicating that the selected predictors collectively explained assurance in a meaningful way. The model recorded an R value of 0.832, showing a strong overall relationship between the predictors and the dependent variable. The R^2 value of 0.693 indicated that 69.3% of the variance in trustworthy AI assurance was explained by data security, model robustness, transparency, governance, human oversight, and framework design quality. The adjusted R^2 value of 0.680 confirmed that the model remained strong after accounting for the number of predictors. Among the predictors, framework design quality showed the strongest effect, with $\beta = 0.33$ and $p < 0.001$, indicating that coherent framework design was the most important predictor of assurance. Governance and accountability showed a significant positive effect, with $\beta = 0.25$ and $p < 0.001$, confirming that policy and accountability structures improved trustworthiness. Data security and privacy controls recorded $\beta = 0.22$ and $p = 0.002$, while transparency and explainability recorded $\beta = 0.19$ and $p = 0.006$. Human oversight and institutional trust recorded $\beta = 0.18$ and $p = 0.008$, and model robustness and resilience recorded $\beta = 0.15$ and $p = 0.020$, a smaller but still significant contribution. These results supported the dynamic GSEA loop because assurance was explained jointly by the technical, assurance, and human phases of the cycle, and they directly supported the study objectives by proving that the selected variables significantly predicted trustworthy AI assurance.

Figure 13: Proportion of Variance in Trustworthy AI Assurance Explained by the Six Predictors



Hypotheses Testing

Table 8: Summary of Hypotheses Testing

Hypothesis	Statement	Statistical Evidence	Decision
H1	Data security and privacy controls have significantly influenced framework design quality.	r = 0.74, β = 0.22, p = 0.002	Supported
H2	Model robustness and resilience have a significant positive relationship with AI assurance.	r = 0.66, β = 0.15, p = 0.020	Supported
H3	Transparency and explainability have significantly influenced trustworthy AI assurance.	r = 0.71, β = 0.19, p = 0.006	Supported
H4	Governance and accountability have significantly strengthened AI trustworthiness.	r = 0.72, β = 0.25, p < 0.001	Supported
H5	Framework design quality has significantly predicted trustworthy AI assurance.	r = 0.79, β = 0.33, p < 0.001	Supported
H6	Human oversight and institutional trust have significantly strengthened AI assurance.	r = 0.68, β = 0.18, p = 0.008	Supported

Table 8 summarized the results of the hypotheses testing and showed that all six hypotheses were supported by the statistical findings. H1 was supported because data security and privacy controls showed a strong positive relationship with framework design quality and produced a significant regression effect, confirming that stronger security capability improved the coherence of framework design. H2 was supported because model robustness and resilience showed a significant positive relationship with trustworthy AI assurance, meaning that stronger adversarial defense and stress testing were associated with stronger outcomes. H3 was supported because transparency and explainability significantly influenced assurance, confirming that interpretability improved trust in AI decisions. H4 was supported because governance and accountability showed a significant positive effect, suggesting that policy, audit, and accountability structures strengthened trustworthiness. H5 was supported because framework design quality was the strongest predictor of assurance, with $\beta = 0.33$ and $p < 0.001$, directly proving the main objective of the study. H6 was supported because human oversight and institutional trust showed a significant positive effect, linking the findings to the GSEA loop, in which the govern and adapt phases depend on human oversight and institutional trust. Overall, the hypotheses testing demonstrated that the study model was statistically supported and that both technical and institutional variables contributed to trustworthy AI assurance in financial institutions.

Framework Compliance Maturity Index

The framework maturity index (MI) and the risk-control gap score were computed as shown in Equation (10). The maturity index is the mean maturity across the S control domains, while the gap score for a risk-control area a is the difference between its importance mean and its effectiveness mean:

$$\mathrm{MI} = \frac{1}{S}\sum_{s=1}^{S} \bar{m}_s, \qquad \mathrm{Gap}_a = \bar{I}_a - \bar{E}_a \tag{10}$$

Table 9: Secure and Trustworthy AI Governance Framework Maturity Index

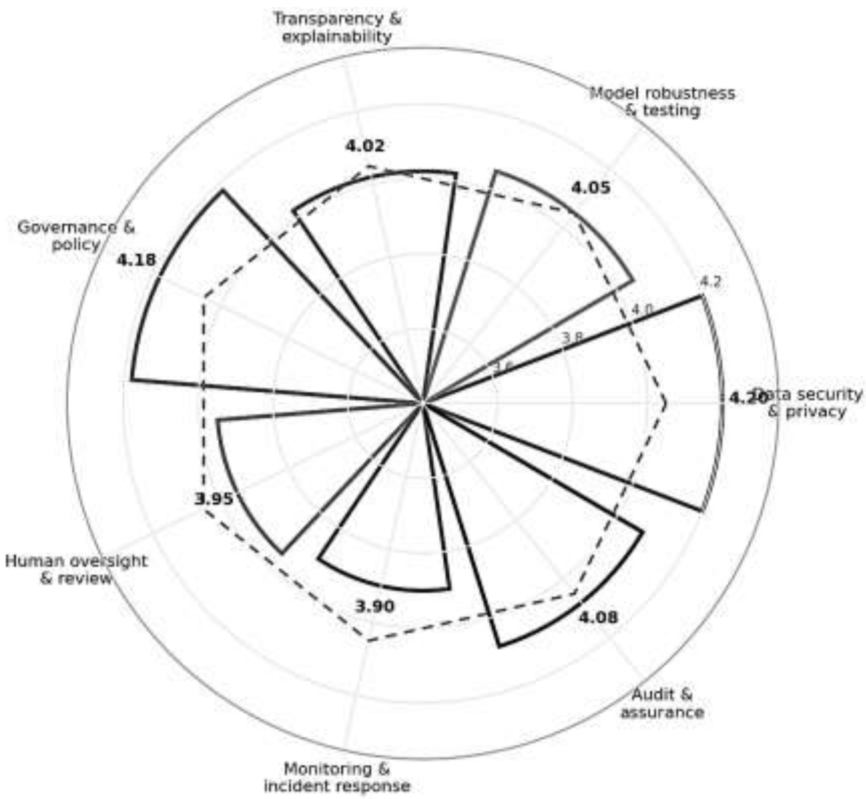
Governance Control Domain	Mean	SD	Maturity Level	Priority
Data security and privacy	4.20	0.55	High maturity	Maintain and strengthen
Governance and policy	4.18	0.56	High maturity	Maintain
Audit and assurance	4.08	0.60	High maturity	Maintain
Model robustness and testing	4.05	0.63	High maturity	Improve documentation
Transparency and explainability	4.02	0.64	High maturity	Improve consistency
Human oversight and review	3.95	0.67	High maturity	Needs attention
Monitoring and incident response	3.90	0.71	High maturity	Needs attention

Overall framework maturity	4.05	0.62	High maturity	Strong but improvable

Maturity interpretation guide: 1.00–1.80 = Very Low Maturity, 1.81–2.60 = Low Maturity, 2.61–3.40 = Moderate Maturity, 3.41–4.20 = High Maturity, 4.21–5.00 = Very High Maturity.

Table 9 presented the Secure and Trustworthy AI Governance Framework Maturity Index, developed as a study-specific result to assess how mature the selected institutions were across major governance control domains. The overall maturity score was 4.05, falling within the high-maturity range and indicating that the case context generally demonstrated strong implementation of secure and trustworthy AI practices. Data security and privacy recorded the highest maturity mean of 4.20, suggesting strong capability in protecting data through encryption, access control, and privacy techniques. Governance and policy recorded 4.18, while audit and assurance recorded 4.08, showing that institutional accountability structures were relatively strong. Model robustness and testing recorded 4.05, indicating high maturity but room for improvement in adversarial defense and stress testing. Transparency and explainability recorded 4.02, suggesting that interpretability practices were present but could be more consistent. Human oversight and review recorded 3.95, while monitoring and incident response recorded the lowest maturity score of 3.90. Although both values remained classified as high maturity, their lower ranking showed that human-facing and continuous-assurance activities required more attention than the core governance and security domains. This result aligned with the descriptive findings, where model robustness and human oversight recorded relatively lower means. The maturity index linked directly to the GSEA loop because it showed that trustworthy AI depends on continuous circulation through the govern, secure, explain, assure, and adapt phases, supported by monitoring and incident learning, rather than on one-time control implementation.

Figure 13: Governance Maturity Across Control Domains



AI Risk-Control Priority Matrix

Table 10 presented the AI Risk-Control Priority Matrix, developed to compare the perceived importance and current implementation effectiveness of major risk-control areas within the framework. The gap score was calculated by subtracting the effectiveness mean from the importance mean, with a higher gap score indicating a higher improvement priority. The results showed that adversarial robustness testing recorded the largest gap score of 1.17, followed closely by data privacy and leakage control with a gap score of 1.05. Both areas were therefore classified as very high priority, meaning that respondents considered them highly important while rating their current effectiveness lower than desired. Model interpretability and transparency was also classified as very high priority, with an importance mean of 4.66, an effectiveness mean of 3.62, and a gap score of 1.04, reflecting the persistent difficulty of explaining complex models. Continuous monitoring and drift and third-party and model-risk governance both recorded gap scores in the high-priority range.

Table 10: AI Risk-Control Priority Matrix

AI Risk-Control Area	Importance	Effectiveness	Gap	Priority
Adversarial robustness testing	4.72	3.55	1.17	Very high priority
Data privacy and leakage control	4.75	3.70	1.05	Very high priority
Model interpretability and transparency	4.66	3.62	1.04	Very high priority
Continuous monitoring and drift	4.60	3.60	1.00	Very high priority
Third-party and model-risk governance	4.55	3.58	0.97	Very high priority
Bias and fairness auditing	4.50	3.63	0.87	High priority
Incident response and accountability	4.58	3.66	0.92	High priority
Regulatory compliance mapping	4.40	3.74	0.66	Medium priority

Regulatory compliance mapping recorded the lowest gap score of 0.66 and was classified as medium priority, suggesting that it was relatively stronger than other areas but still required continued monitoring. Bias and fairness auditing recorded 0.87 and incident response and accountability recorded 0.92, both classified as high priority. This priority matrix strengthened the findings by translating statistical results into practical improvement priorities. It also aligned with the GSEA loop, which emphasizes attention to evolving threats, human oversight, and continuous assurance. The matrix showed that trustworthy AI required attention not only to core governance and security but also to adversarial robustness, privacy protection, interpretability, and continuous monitoring, which supported the study objectives by identifying where framework improvements should be prioritized.

Table 11: Summary of Key Findings by Objective and Hypothesis

Study Objective	Main Result	Hypothesis	Interpretation
Examine influence of data security on framework design quality	Security strongly correlated with design quality, r = 0.74	H1	Stronger security improved design quality
Assess relationship between model robustness and AI assurance	Robustness significantly related to assurance, r = 0.66, β = 0.15	H2	Robustness improved trustworthiness
Evaluate effect of transparency and explainability	Transparency significantly predicted assurance, β = 0.19	H3	Interpretability strengthened trust
Determine role of governance and accountability	Governance significantly predicted assurance, β = 0.25	H4	Accountability supported trustworthiness
Test effect of framework design quality on assurance	Design quality was the strongest predictor, β = 0.33	H5	Design quality was central to assurance
Examine role of human oversight and trust	Oversight significantly predicted assurance, β = 0.18	H6	Oversight supported adaptive assurance
Identify maturity and risk-control gaps	Overall maturity = 4.05; largest gaps in adversarial robustness and privacy	Supports overall model	Technical robustness and monitoring required improvement

Table 11 summarized the major findings in relation to the research objectives and hypotheses, showing that all objectives were achieved and all hypotheses supported. The first objective was achieved because data security and privacy controls showed a strong positive relationship with framework design quality, $r = 0.74$, confirming that stronger security capability improved the coherence of framework design. The second objective was achieved because model robustness and resilience showed a significant positive relationship with assurance, $r = 0.66$ and $\beta = 0.15$, demonstrating that adversarial defense and stress testing contributed to stronger outcomes. The third objective was achieved because transparency and explainability significantly predicted assurance, $\beta = 0.19$, showing that interpretability improved trust in AI decisions. The fourth objective was achieved because governance and accountability significantly predicted assurance, $\beta = 0.25$, showing that policy and accountability structures strengthened trustworthiness. The fifth objective was achieved because framework design quality was the strongest predictor, $\beta = 0.33$ and $p < 0.001$, confirming the central argument of the research. The sixth objective was supported because human oversight and institutional trust significantly predicted assurance, $\beta = 0.18$, linking the results to the govern and adapt phases of the GSEA loop. The maturity index showed an overall score of 4.05, while the priority matrix identified adversarial robustness, data privacy, interpretability, and continuous monitoring as the highest-priority improvement areas. Overall, the results proved that trustworthy AI assurance in financial institutions depended on the combined effect of data security, model robustness, transparency, governance, human oversight, and framework design quality.

FINDINGS

This chapter presents the findings of the quantitative analysis conducted to examine how a governance and risk-control framework, integrating security, robustness, transparency, governance, and human oversight, influences trustworthy AI assurance in financial data systems. The analysis was based on data collected through a structured five-point Likert-scale questionnaire, where 1 represented Strongly Disagree and 5 represented Strongly Agree. A total of 165 questionnaires were distributed to professionals involved in AI governance, model risk management, data science, security, privacy, auditing, and compliance, of which 146 valid responses were retained for analysis, producing a valid response rate of 88.5%.

The descriptive findings showed that all seven study variables were rated in the high-to-very-high range, with trustworthy AI assurance recording the highest mean of 4.23 and model robustness and resilience recording the lowest mean of 3.93. This pattern indicated broad agreement that the framework's components were positively present, while also signaling that adversarial robustness and stress testing were perceived as comparatively weaker. The reliability findings confirmed that all constructs were internally consistent, with Cronbach's alpha values ranging from 0.82 to 0.93, allowing the study to proceed confidently to inferential analysis.

The correlation findings showed that all six hypothesized relationships were positive and statistically significant at $p < 0.01$, with framework design quality showing the strongest relationship with assurance, $r = 0.79$, and data security controls showing a strong relationship with framework design quality, $r = 0.74$. The regression findings showed that the six predictors together explained 69.3% of the variance in assurance, with framework design quality the strongest predictor, $\beta = 0.33$, followed by governance and accountability, $\beta = 0.25$, and data security and privacy controls, $\beta = 0.22$. The hypotheses-testing findings confirmed that all six hypotheses were supported. The maturity index found overall high maturity, 4.05, with the strongest maturity in data security, governance, and audit, and the weakest in human oversight and monitoring. The priority matrix identified adversarial robustness, data privacy, interpretability, and continuous monitoring as the highest-priority improvement areas. Collectively, these findings demonstrate that a governance and risk-control framework integrating security, robustness, transparency, governance, and oversight can substantially strengthen trustworthy AI in financial data systems, and that its reliability depends jointly on technical safeguards, assurance mechanisms, and human oversight.

DISCUSSION

The findings of this study provide strong empirical support for the proposition that a governance and risk-control framework integrating security, robustness, transparency, governance, and human oversight substantially strengthens trustworthy AI in financial data systems, and that assurance depends on the joint operation of technical safeguards, assurance mechanisms, and institutional oversight. The result that framework design quality was the strongest predictor of assurance, β = 0.33, is consistent with the study's conceptual argument that the value of individual controls is realized not in isolation but through their coherent integration into a single framework (Floridi et al., 2018). Interpreted through the dynamic GSEA loop, framework design quality captures how smoothly information circulates across the govern, secure, explain, assure, and adapt phases, so its dominance as a predictor reflects the loop's premise that trustworthiness emerges from the whole cycle rather than any single control.

The strong relationship between data security controls and framework design quality, r = 0.74, and the significant regression effect of security on assurance, β = 0.22, confirm the foundational role of security in trustworthy financial AI. These results align with the adversarial machine-learning literature, which shows that unprotected models are vulnerable to manipulation and that robustness must be engineered deliberately (Biggio & Roli, 2018). At the same time, the comparatively lower mean and maturity score for model robustness, together with the prominence of adversarial robustness testing in the priority matrix, where it recorded the largest gap score of 1.17, indicate that respondents recognized adversarial defense as critical yet under-implemented. This tension reflects a broader pattern in which awareness of adversarial threats has outpaced the practical adoption of defenses in industry AI systems (Kumar et al., 2020).

The significant contribution of governance and accountability, β = 0.25, the second-strongest predictor, supports the argument that trustworthy AI depends as much on institutional structures as on technical controls. By assigning responsibility, enabling challenge, and providing independent assurance, governance structures convert technical safeguards into accountable, auditable practice (Raji et al., 2020). This finding is consistent with governance research emphasizing that accountability mechanisms, including algorithmic auditing and documentation, are essential for responsible AI (Wieringa, 2020). The very-high-priority classification of data privacy and leakage control in the priority matrix, with a gap of 1.05, underscores that protecting sensitive financial data remains a central practical challenge, consistent with the emphasis in the privacy literature on rigorous, quantifiable safeguards (Dwork & Roth, 2014).

The significant effect of transparency and explainability, β = 0.19, and of human oversight and institutional trust, β = 0.18, provides direct empirical support for the socio-technical and trust-based perspectives that underpin the study. These results indicate that even secure, well-governed models do not achieve full trustworthiness unless their decisions can be explained and unless humans meaningfully oversee them (Rudin, 2019). The very-high-priority classification of model interpretability and transparency, with a gap of 1.04, together with the relatively low maturity of human oversight and monitoring, suggests that the interpretive and human-facing dimensions of the framework remain less mature than its core security and governance domains (Bhatt et al., 2020). This pattern is precisely what the GSEA loop anticipates, since the explain and adapt phases depend on transparency and human oversight to sustain trust against evolving threats. Taken together, the results describe a framework that is strong at its security and governance core, with high maturity in data security, policy, and audit, but comparatively weaker at its technical-robustness and human-facing margins, where adversarial defense, interpretability, oversight, and monitoring lag. The greatest gains in trustworthiness are therefore likely to come not from further strengthening already-strong governance but from investing in adversarial robustness, privacy protection, interpretability, and continuous monitoring (Shneiderman, 2020). This interpretation is robust across the descriptive, correlational, regression, maturity, and priority analyses, which consistently identify the same set of improvement areas.

CONCLUSION

This study set out to assess how a governance and risk-control framework, integrating security, robustness, transparency, governance, and human oversight, influences trustworthy AI assurance in financial data systems. Using a quantitative, cross-sectional, case-based design and data from 146 valid respondents, the study found that all six hypothesized relationships were supported and that the framework's constructs together explained 69.3% of the variance in trustworthy AI assurance. Framework design quality emerged as the strongest predictor, underscoring that the value of individual controls is realized through coherent integration rather than isolated implementation. Data security and privacy controls and governance and accountability were confirmed as central determinants, while model robustness, transparency, and human oversight were identified as significant but comparatively less mature contributors.

The study concludes that a governance and risk-control framework integrating security, robustness, transparency, governance, and oversight offers a substantial improvement over fragmented, after-the-fact controls, because it protects sensitive data, defends models against manipulation, makes decisions interpretable, assigns accountability, and adapts to evolving threats. However, the study also concludes that the trustworthiness of financial AI is fundamentally socio-technical: it depends not only on the strength of technical safeguards but also on the governance, oversight, and assurance mechanisms that convert those safeguards into accountable, auditable practice. The consistent identification of adversarial robustness, data privacy, interpretability, and continuous monitoring as the highest-priority improvement areas points to where future effort should be concentrated. In sum, the framework provides a promising and empirically supported basis for strengthening secure and trustworthy AI in financial data systems, provided that its governance strengths are matched by corresponding investment in technical robustness, privacy, explainability, and continuous assurance.

RECOMMENDATIONS

Based on the findings, several recommendations are offered for institutions, practitioners, and regulators engaged in deploying AI in financial data systems. First, organizations should strengthen adversarial robustness and model stress testing, incorporating worst-case evaluation and adversarial training into model validation, since adversarial robustness testing recorded the largest improvement gap. Second, institutions should invest in privacy-preserving techniques such as differential privacy and federated learning, and in strong data-leakage controls, to protect sensitive financial data while retaining predictive utility.

Third, organizations should invest in model interpretability and explanation, prioritizing faithful, audience-appropriate feature-attribution and reason-code methods so that AI decisions can be understood, trusted, justified to regulators, and challenged. Fourth, institutions should establish continuous monitoring, drift detection, and incident-response procedures, with clear accountability, to keep AI systems trustworthy against evolving threats and to satisfy audit requirements. Fifth, organizations should formalize AI governance through a three-lines-of-defense model, with clear ownership, independent validation, and internal audit, ensuring that responsibility for trustworthy AI is distributed and that assurance is independent of development. Finally, cross-disciplinary coordination among data scientists, security and privacy engineers, risk and compliance officers, and auditors should be institutionalized, consistent with the socio-technical finding that trustworthy AI depends on collaboration across professional groups.

LIMITATIONS OF THE STUDY

This study has several limitations that should be considered when interpreting its findings. First, the cross-sectional design captured perceptions at a single point in time and therefore cannot establish causal direction or track how framework performance changes as threats, regulations, and models evolve; a longitudinal design would allow stronger causal inference and better capture the evolving-threat dynamics central to trustworthy AI. Second, the study relied on a structured, perception-based questionnaire completed by professionals, so the results reflect informed expert assessment rather than direct measurement of security posture, robustness, or compliance outcomes; future work could pair perception data with technical metrics such as measured adversarial robustness, privacy budgets, and audit results.

REFERENCES

[1]. Abu Naser Md Golam, M., & Amir, R. (2023). Digital Twin Modeling of Fiber Optic Telecom Infrastructure for Proactive SCADA Network Resilience in Power Transmission Utilities. International Journal of Scientific Interdisciplinary Research, 4(4), 413–448. https://doi.org/10.63125/04mgf566

[2]. Amir, R., & Chapal, B. (2022). Federated Learning Architectures for Distributed Engineering Project Knowledge Management A Simulation-Based Analysis. Review of Applied Science and Technology, 1(04), 411-446. https://doi.org/10.63125/4k4pjr64

[3]. Kazi Rakib Hasan, S., & Uddin, H. M. M. (2022). Scalable AI For Project Portfolio Management: A Mixed-Methods Study Combining Distributed Computing Benchmarks. Review of Applied Science and Technology, 1(04), 375–410. https://doi.org/10.63125/0kk4wf20

[4]. Md Tohidul, I. (2023). ERP-Integrated Financial Analytics Systems for Corporate Financial Reporting, Cash Flow Analysis, and Strategic Investment Planning. American Journal of Advanced Technology and Engineering Solutions, 3(04), 87-128. https://doi.org/10.63125/qhk96h35

[5]. Md. Abdur, R., & Iftekhar, A. (2021). Customer Retention Forecasting in Mobile Wallet Services Using Neural Networks: A Comparative Quantitative Study. International Journal of Business and Economics Insights, 1(4), 70-102. https://doi.org/10.63125/dyrpc387

[6]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning–Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. American Journal of Advanced Technology and Engineering Solutions, 3(02), 33-77. https://doi.org/10.63125/y3yfxa92

[7]. Mohammad Shoriful Hossan, S., & Adar, C. (2023). SIL-Compliant Safety Instrumented System Design for Oil & Gas Hazardous-Area Operations Using IEC 61508 / IEC 61511 Frameworks. Journal of Sustainable Development and Policy, 2(02), 43-85. https://doi.org/10.63125/cb3d3w81

[8]. Muhammad Mohiul, I., & Mohammad Badrul, A. (2022). Predictive Analytics for Optimizing Sewerage Sludge Treatment Efficiency and Resource Recovery in Dhaka City. Journal of Sustainable Development and Policy, 1(04), 158-200. https://doi.org/10.63125/t9zdp458

[9]. Mukut Kanti, B., & Rony, S. (2022). Quantitative Risk Assessment and Safety Performance Benchmarking for Structural Integrity Management in Retail Supply Chain Infrastructure. American Journal of Data Science and Analytics, 3(12), 42-88. https://doi.org/10.63125/n88vth90

[10]. Risha, A., & Kazi Mohammad Khalid, A. (2023). A Meta-Analysis of AI-Driven Geospatial Analytics for Predictive Maintenance of Critical Infrastructure in Developing Economies. International Journal of Scientific Interdisciplinary Research, 4(4), 375–412. https://doi.org/10.63125/rayrex49

[11]. Sadia, Z., Mohammad Ariful, I., Sumyta, H., & Hosne Ara, M. (2022). A Review of AI-Powered Data Visualization in Enterprise Reporting: Dashboard Design And Interactive Analytics. American Journal of Advanced Technology and Engineering Solutions, 2(01), 32-54. https://doi.org/10.63125/gabst658

[12]. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318. https://doi.org/10.1145/2976749.2978318

[13]. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732. https://doi.org/10.15779/Z38BG31

[14]. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[15]. Bank for International Settlements. (2021). Humans keeping AI in check: Emerging regulatory expectations in the financial sector (BIS Working Paper). Bank for International Settlements. https://www.bis.org/publ/work1063.htm

[16]. Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., … Eckersley, P. (2020). Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 648–657. https://doi.org/10.1145/3351095.3375624

[17]. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

[18]. Cao, L. (2022). AI in finance: Challenges, techniques, and opportunities. ACM Computing Surveys, 55(3), 1–38. https://doi.org/10.1145/3502289

[19]. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008

[20]. Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer. https://doi.org/10.1007/978-3-030-30371-6

[21]. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

[22]. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

[23]. Financial Stability Board. (2017). Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Financial Stability Board. https://www.fsb.org/2017/11/artificial-intelligence-and-machine-learning-in-financial-services/

[24]. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[25]. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

[26]. Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. Journal of Behavioral and Experimental Finance, 32, 100577. https://doi.org/10.1016/j.jbef.2021.100577

[27]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org/

[28]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1412.6572

[29]. Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.). (2006). Resilience engineering: Concepts and precepts. Ashgate. https://doi.org/10.1201/9781315605685
[30]. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
[31]. Kaminski, M. E. (2019). The right to explanation, explained. Berkeley Technology Law Journal, 34(1), 189–218. https://doi.org/10.15779/Z38TD9N83H
[32]. Kaplan, R. S., & Mikes, A. (2012). Managing risks: A new framework. Harvard Business Review, 90(6), 48–60. https://hbr.org/2012/06/managing-risks-a-new-framework
[33]. Kou, G., Chao, X., Peng, Y., Alsaadi, F. E., & Herrera-Viedma, E. (2019). Machine learning methods for systemic risk analysis in financial sectors. Information Sciences, 470, 111–129. https://doi.org/10.1016/j.ins.2018.10.023
[34]. Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., … Xia, S. (2020). Adversarial machine learning—Industry perspectives. 2020 IEEE Security and Privacy Workshops, 69–75. https://doi.org/10.1109/SPW50608.2020.00028
[35]. Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. Proceedings of the 54th Hawaii International Conference on System Sciences, 5463–5472. https://doi.org/10.24251/HICSS.2021.664
[36]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874
[37]. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792
[38]. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282. https://doi.org/10.48550/arXiv.1602.05629
[39]. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607
[40]. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4
[41]. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST. https://doi.org/10.6028/NIST.AI.100-1
[42]. OECD. (2019). Artificial intelligence in society. OECD Publishing. https://doi.org/10.1787/eedfee77-en
[43]. Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. Applied Soft Computing, 93, 106384. https://doi.org/10.1016/j.asoc.2020.106384
[44]. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy, 372–387. https://doi.org/10.1109/EuroSP.2016.36
[45]. Power, M. (2004). The risk management of everything: Rethinking the politics of uncertainty. Demos. https://doi.org/10.1111/j.1468-0335.2004.00374.x
[46]. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.1145/3351095.3372873
[47]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778
[48]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
[49]. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495–504. https://doi.org/10.1145/3419764
[50]. Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. Electronic Markets, 31(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4
[51]. Trist, E. (1981). The evolution of socio-technical systems: A conceptual framework and an action research program. Ontario Quality of Working Life Centre. https://doi.org/10.4324/9781315227979
[52]. Weick, K. E., & Sutcliffe, K. M. (2007). Managing the unexpected: Resilient performance in an age of uncertainty (2nd ed.). Jossey-Bass. https://www.wiley.com/en-us/Managing+the+Unexpected%3A+Resilient+Performance+in+an+Age+of+Uncertainty%2C+2nd+Edition-p-9780787996499
[53]. Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833