AI-Enabled Financial Integrity Engines: Explainable Models for Transparent Risk Assessment and Human-Centered Oversight

Mariana Tataryn[1]

[1]Deloitte Inc.
San Francisco, California, United States

***Abstract*— As financial systems get increasingly digitized, organizations are encountering increasingly high risks of algorithmic opacity, regulatory non-compliance, auditability gaps, and loss of institutional trust. Though artificial intelligence is now a mainstream instrument used in the assessment of financial risk, anomaly detection, and predictive analytics, its fast mainstreaming has also revealed structural vulnerabilities in governance arch design that is based on black-box models and automation-driven logic. These are the problems that highlight the increased significance of financial integrity systems that can entrench explainability, regulatory compliance, and human-centered oversight, as opposed to technology-based AI deployment strategies. This paper takes a governance-based approach to study the transformations occurring in transparency, risk containment, and integrity results in financial systems with the help of AI-enabled Financial Integrity Engines. The study, based on institutional economics and explainable AI theory, takes AI not as an independent decision-maker, but as an embedded governance mechanism, and the quality of which is determined by explainability, compliance-by-design, and human-in-the-loop control mechanisms. The empirical test is followed by the use of secondary longitudinal panel data of five developed economies (2020-2024). In the study, the fixed-effects econometric model is utilized to determine the effect of AI adoption intensity, explainable AI, embedded compliance capacity and human-centered oversight on a composite Financial integrity Index. The interaction effects are also included in order to reflect the conditionality of AI efficacy in varying governance set-ups. The findings prove that AI implementation does not produce significant financial integrity outcomes. Conversely, explainable AI and embedded compliance systems have significant and significant positive effects that are statistically significant in every environment under observation. The relationship between the adoption of AI and explainability enhances the impact of integrity by 35-55 percent, and human oversight of the AI process increases the impact of risk mitigation and transparency. The most integrity gains are observed in countries with a developed compliance architecture and organized human oversight framework, but decreasing returns are observed in technology-intensive settings with too little explainability. The results suggest that the results of financial integrity depend not so much on AI predictive accuracy or intensity but on governance architecture. AI systems can help in managing risks sustainably only when integrated as part of transparent, audited and human controlled institutional structures. This paper concludes that the explainable nature and oversight are not ancillary aspects of responsible AI application in finance, but among the requisites. This framework should be expanded by future research investigating firms and comparing them in the context of new financial systems.**

Keywords— financial integrity; explainable artificial intelligence; AI governance; risk assessment; compliance-by-design; human-in-the-loop; financial transparency; institutional oversight; fintech regulation.

I. INTRODUCTION

Financial systems today are operating in a highly complex and unstable institutional environment of fast-paced digitalization, mounting regulatory demands, intensified data volumes and increased systemic and non-systemic risk exposure. Financial decision-making is not limited to periodic reporting, and retrospective control anymore; it has developed to become a continuous risk anticipation, compliance assurance, and real-time governance across interrelational organizational and market structures. Here, financial information and risk assessment mechanisms are what will determine the stability of the economic system, institutional trust, and value creation in the long term.

Rapid replacing of artificial intelligence (AI) with financial management has profoundly redefined the structure of risk identification, predicting and managing. The use of AI-based tools is becoming common to detect anomalies, forecast financial distress, and automate reconciliations and assist in

ASEJ - Scientific Journal of Bielsko-Biala School of Finance and Law
Volume 29, No 4 (2025), pages 9
https://doi.org/10.19192/wsfip.sj4.2025.15
Received: September 2025 Accepted: December 2025
Published: December 2025

making strategic decisions. Although these technologies hold efficiency benefits and increased levels of analysis, they are also being faced with new governance issues regarding the aspects of transparency, responsibility and legitimacy of decisions. Control, explainability and human responsibility are fundamental in the delegation of judgment to algorithmic systems in finance where the regulatory scrutiny, auditability, and fiduciary responsibility are central.

Although the use of AI applications in financial processes has increased, the leading implementation practices are still largely technology-oriented. AI systems are often presented as performance-enhancing systems that are aimed at speed in automation, predictive accuracy, or cost reduction but little thought is given to the institutional embedding of these systems. Consequently, a lot of AI-based financial products are provided as black boxes, and their results are hard to interpret, justify, or audit. These shapes can make operations work better in the short-term, but tend to compromise transparency, weaken compliance guarantees and erode user trust - especially in regulated settings where explainability and traceability are non-negotiable conditions.

The continued adoption of the technology-first form of AI indicates the presence of a structural gap in the practice of finance, as well as in academic research. Although the current body of literature has exhaustively explored the issue of AI accuracy, algorithmic performance, and digital transformation effects, much less focus has been given to AI as a mechanism of institutional governance that determines financial integrity. Existing research is more likely to study AI tools individually and not in a systematic manner that considers the impact that explainability, compliance architecture, and human oversight have on its effectiveness. Therefore, the outcomes of financial risk are usually traced to technological sophistication instead of governance structures overall that AI systems are implemented in.

The research problem developed in the present study is caused by the absence of an analytical framework that can be fully comprehensive to define how AI-enabled financial systems can affect transparency, risk containment, and integrity under different institutional settings. Specifically, there exist few empirical studies on whether the use of AI can improve financial integrity on its own or whether it fundamentally requires complementary governance initiatives (including explainable models, embedded compliance, and human-centered oversight). This discontinuity limits the capacity of policymakers, regulators and financial executives to create AI systems that cannot only be efficient, but also reliable, auditable, and trustworthy.

The aim of the current paper is to evaluate AI-based financial integrity engines in the governance-oriented perspective by prioritizing explainability, institutional adherence, and the human-in-the-loop control as the fundamental conditions enabling it. The vision of the study is the conceptualization of financial AI as a supportive infrastructure to promote integrity and is integrated into the ordered organizational systems. The research examines the influence of various combinations of AI adoption, elucidate modelling, capacity to comply and human control on the outcomes of financial integrity in the long run by empirically reviewing cross-country panel data.

The study will have the following objectives:

- to determine how the financial integrity outcomes directly depend on the adoption of AI;
- to study how explainable AI can contribute to increasing transparency and risk accountability;
- to determine the extent to which human-centered regulation moderates the performance of AI-based financial systems;
- to examine how embedded mechanisms of compliance contribute to risk management using AI;
- to establish a unified empirical approach under which AI is the element of financial integrity contingent upon governance and not an independent technological answer.

The originality of this study is that it combines the financial governance theory, explainable artificial intelligence, and institutional economics under the umbrella analytical framework of financial integrity. Unlike the literature which focuses on the performance of algorithms or the efficiency of digital procedures individually, this paper redefines financial integrity as a resultant process of interactions between technology, regulation, and human judgment. It moves the concept of AI-powered Financial Integrity Engines, where explainability and control are not side effects of AI deployment in finance, but a feature of AI deployment in finance.

These theoretical and practical implications of this research fall across various areas. In the case of financial economics and FinTech research, it develops governance-conscious AI as a key analysis type of risk management and compliance research. In the case of the institutional theory, it emphasizes the importance of explainability and control in stabilizing decision systems that are based on algorithms. The results of the study remind practitioners and regulators that successful implementation of AI in finance cannot be realized by the excessive automation of financial systems, but rather through the careful design of transparent, responsible, and human-oriented financial intelligence systems. In the end, the positions of the study, which explainable and controlled AI is a key mechanism through which the financial transparency, regulatory trust, and economic resilience will be mutually achieved in the long term.

II. LITERATURE REVIEW

The accelerated spread of artificial intelligence to the financial system has aggravated the interest of academics toward the transparency, accountability, and control of algorithmic decision-making. An emerging literature is converging on the assumption that although AI improves the analytical ability and efficiency in operations, unregulated or non-transparent use can compromise trust, compliance, as well as institutional legitimacy. In controlled systems like finance, the question is no longer one of whether AI should be implemented, but on what terms of governance it can play a role in creating a sustainable financial integrity rather than enhance systemic risk.

Another concept that has come to the fore of this debate is

explainable artificial intelligence (XAI). In their review of insurance applications, Owens et al. (2022) reveal that model explainability is vital both in regulatory compliance and internal decision validation and stakeholder trust. Their results indicate that predictive accuracy by itself is inadequate in high stakes financial conditions, instead, explainability acts as a control mechanism that allows auditability and responsibility attribution. This observation is reflected in financial fraud detection literature, as the studies by Aljunaid et al. (2025) demonstrate that explainable AI structures can be quite helpful in improving transparency and reliability of the banking system when used in conjunction with secure learning paradigms. Combined, these works establish the XAI as an integrity-oriented financial AI structural requirement and not an extra technical characteristic.

In addition to explainability, other researchers also underline the relevance of process-based AI governance. Hohma and Lütge (2023) state that a reliable AI cannot be developed based on abstract ethical concepts only but that the concept of trustworthiness needs to be integrated throughout the lifecycle of AI development. Their paradigm changes the focus on the result-based evaluations to the procedures of governance and emphasize the documentation, accountability, and human control as the main aspects. This view corresponds to the risk-aware value creating strategy offered by Ricciardi Celsi (2023), who uses the ideas of AI governance as a critical balancing system between values of innovation and regulatory standards. The contributions of both provide a degree of emphasis that AI systems gain legitimacy not due to technical complexity, but the institutional frameworks that regulate the implementation of the systems.

Another crucial aspect of the AI-enabling financial systems is fairness and bias. Chen et al. (2023) present an important overview of the issues of fairness in data management and analytics and show how discriminatory results can be promoted through biased data pipelines and obscure model reasoning. Their discussion supports the thesis that fairness can only be enforced by transparency and explainability. In line with this perspective, Yaseen and Al-Amarneh (2025) provide strong empirical evidence that the trust in AI-based fraud detection in banking is highly mediated by the perception of transparency and fairness, but not by the performance indicators. All these findings positively indicate that fairness is not a standalone issue of ethical consideration, but rather, it is a part of financial integrity and effectiveness of governance.

Human-centered oversight is seen as a stabilizing in AI governance systems and its role is getting more and more significant. Seralidou et al. (2025) propose a human-based trustworthiness risk evaluation model (AI_TAF), which explicitly incorporates the human judgment in the process of evaluating the risk of AI systems. Their output proves that the notion of trustworthiness comes up as a result of interaction between algorithmic outputs and human interpretation, especially during uncertain and challenging situations of decision making. The same findings can be made in studies on healthcare-oriented algorithms relying on algorethics, where Lastrucci et al. (2024) claim that the erosion of integrity occurs as a result of innovation, even in technologically advanced algorithms, without being controlled by a structured human. Though not directly in finance, their observations can be directly applied into the field of financial governance where human accountability is legally and ethically unavoidable.

The recent research also broadens the debate to include the financial transparency and corporate governance. The article by Shanab and Omoush (2025) offers empirical data on the Jordanian context that demonstrates that AI-based accounting and reporting systems will improve the quality of transparency and governance with the implementation of institutional control. They however warn that unless automation is regulated and professionally supervised, it will tend to blur rather than clarify financial data. In the same manner, Choowan and Daovisan (2026) in their systematic review of AI in data governance to make financial decisions find that, it is governance maturity and not AI intensity, which dictates whether adoption of AI would enhance decision quality and risk management. Their creation supports the thesis that AI is an enabler that depends on governance instead of being an independent solution.

Combined, the literature brings up a number of overlapping insights. To start with, explainability is always a requirement of trust, auditability and compliance to financial AI systems. Second, anthropocentric control is the only way to put algorithms into context and ensure accountability. Third, the relationship between AI adoption and integrity outcomes is mediated by governance architecture, which includes development processes, compliance integration, and fairness controls. Nonetheless, in spite of these developments, current literature is still scattered across various fields like insurance, banking, healthcare, and data governance; without a coherent empirical framework that combines explainability, oversight, and compliance into one analytical framework.

This paper fills this gap by summarizing the results of explainable AI, institutional governance and financial integrity literature in a unified empirical construct. The current study expands on the current body of literature by conceptualizing AI-enabled Financial Integrity Engines as governance-integrated systems and not as a technology per se and presents cross-country econometric data on the presence of explainability and human control in the appropriate conditioning of AI effectiveness in financial risk management and transparency.

III. MATERIALS AND METHODS

A. Research design

The research design taken in this study is a quantitative and explanatory study, which has an objective of identifying and evaluating the institutional circumstances in which AI-enabled financial systems can play a role to enhance financial integrity and risk transparency. The study is based on a governance-located analytical scheme, where artificial intelligence is developed as an incorporated image of financial control systems, as opposed to a decision-making device.

The research uses a panel data model to address cross-

country heterogeneity and time dynamics in the study. This design will allow examining how changes in AI adoption, explainability, capacity to comply with and human oversight impact financial integrity outcomes over time. The longitudinal design is especially appropriate when it comes to exploring how AI governance mechanisms are gradually getting institutionalized and what their cumulative impacts will be.

The empirical strategy is concerned with determining the conditional and interaction effects, which represents the main theoretical hypothesis according to which AI performance in finance becomes conditional upon the explainability and the oversight of a human being. The fixed-effects estimation is applied to regulate the unobserved time-invariant institutional features, whereas time effects embrace the global shocks and macro-financial tendencies.

B. Selection of samples and time of observation.

The empirical sample is made up of five developed economies, including United States, Germany, France, Japan, and United Kingdom. The selection of these countries was done on three grounds:

1) high rates of AI usage on financial and regulatory processes;
2) well-developed and documented financial governance and compliance structures;
3) cross-country data on the governance of AI, financial integrity and institutional quality is available and consistent.

The time frame of observation is 2020-2024 which will capture the boost in the use of AI in the financial sector after the shock of COVID-19 and the consequent maturation in regulatory and governance reactions. The given period is especially pertinent in regards to evaluating the way in which crisis-induced digitalization transformed into even more organized AI integration that is more governance conscious. The last data is a balanced panel making it comparable across the countries and years which helps in making strong econometric inference.

C. Sources of data and data collection methods.

To achieve transparency, replicability and methodological rigor, the study will solely use secondary data which has been collected using internationally acknowledged and publicly available sources.

Key data sources include:

1) AI adoption and digitalization international databases (e.g., OECD, World Bank);
2) governance, quality of regulations and rule of law indicators;
3) financial risk management and indices with regard to integrity;
4) variables of macroeconomic and financial control based on the official statistical repositories.

Financial Integrity Index is a dependent variable that has been formulated as a composite variable or measure of transparency, internal control effectiveness, and risk management performance. The independent variables imply the intensive use of AI, elucidation and visibility of AI systems, ability to have human control over the systems and inherent quality of compliance. Where necessary, all the variables were normalized so as to make cross-country comparison a possibility. The consistency of data was checked to detect the absence of values, the presence of outliers and structural discontinuities. In isolated missing observations where a linear interpolation was used, overall trend dynamics were not affected.

D. Econometric model

In order to test the hypotheses of the research empirically, the following fixed-effects panel regression model is estimated:

$$FIit = \alpha + \beta 1AIit + \beta 2XAIit + \beta 3HCOit + \beta 4COMPit + \beta 5(AIit \times XAIit) + \beta 6(AIit \times HCOit) + \gamma Xit + \mu i + \lambda t + \varepsilon it \quad (1)$$

where:

- FIit - denotes the Financial Integrity Index for country i in year t;
- AIit - κepresents AI adoption intensity in financial processes;
- XAIit- captures the degree of explainability and transparency of AI systems;
- HCOit - reflects human-centered oversight capacity;
- COMPit - denotes embedded compliance and regulatory quality;
- Xit - is a vector of control variables;
- μi and λt - represent country and time fixed effects;
- εit - is the error term.

Under this model, α indicates the baseline level of financial integrity when the explanatory variables equal zero and β 1 0-4 indicates the direct marginal impact of the adoption of AI, explainable AI, human-centered oversight, and embedded compliance on financial integrity.

Coefficients, 5, and 6 reflects the effect of interaction: How the influence of AI adoption on financial integrity varies with the existence of explainability and human-centered oversight, respectively.

By incorporating the term of interaction, the analysis will be able to generate the effect of conditionality; the purpose of directly testing whether AI adoption positively influences financial integrity is conditionalized by explainability and human supervision. Large standard errors are used to explain heteroskedasticity and within-panel correlation.

Hypotheses:

H1: The positive effects of AI use on financial integrity can only be provided when the use is explicable.

H2: Explainable AI mediates the correlation between AI adoption and reduction of risk.

H3: Human-centered monitoring enhances the efficiency of AI-based financial integrity engines.

H4: AI positively impacts the financial transparency, which is enhanced by the presence of embedded compliance mechanisms.

E. Validation and reliability

A number of operations are implemented to render validity and reliability of the empirical findings. To begin with, the diagnostics of multicollinearity shows that there is reasonable variance inflation, which implies that the estimated coefficients are consistent and can be interpreted. Second, also the qualitatively consistent results obtained by alternative model specifications that have altered control sets confirm the robustness. Third, within-panel explanatory power (R2) does not change with specifications, which has a strong signal of good model performance. The temporal fixed effects manage the world shocks and common trends, and the country fixed effect controls the bias of the presence of unobserved institutional heterogeneity. External validity is even further enhanced by the similarity of signs and the level of significance of the coefficients around the world.

F. Limitations

In spite of its strengths, it has been linked to a number of limitations. To start with, the aggregate country-level measures can conceal heterogeneity at the firm level in the adoption and governance of AI. Second, because panel estimation addresses the endogeneity issues, causal inference is limited by observational characteristics of the data. Third, the simplification and weighting assumptions required in the construction of composite indices are bound to affect the absolute coefficient magnitude.

Lastly, the concentrations of the study in the advanced economies restricts the ability to generalize the findings to emerging or developing financial systems with different institutional conditions and data availability. These weaknesses indicate research directions in the future such as company-level studies and greater geographical scale.

IV. RESULTS

A. Model description and estimation logic

The empirical study relies on the balanced panel data starting in 2020 and ending in 2024 in the United States, Germany, France, Japan, and the United Kingdom. The model aims at evaluating the impact of AI-powered Financial Integrity Engines, together with explainability and human-centric oversight systems, on financial risk transparency and integrity outcomes.

The dependent variable is a composite Financial Integrity Index (FI) that reflects the measurement of risk management, high-quality internal controls, and outcomes of transparency. The main factors of explanation are AI Adoption (AI), Explainable AI (XAI), Human-Centered Oversight (HCO), and Embedded Compliance Capacity (COMP). The terms of interaction (AI × XAI and AI × HCO) are implemented to determine whether the performance of AI is conditional on the situation of governance and oversight.

The fixed-effects panel regression on country and time effects estimated the model, controlling due to the macroeconomic development, the depth of financial markets, and the intensity of digitalization. Strong standard errors are used to solve the heteroskedasticity and within-panel correlation.

B. Aggregate regression findings.

The results on the estimated coefficients in each country were provided in Table 1. In each of the five economies, the findings indicate a stable and statistically significant association between variables on AI and the financial integrity outcomes.

TABLE 1. PANEL REGRESSION RESULTS: AI-ENABLED FINANCIAL INTEGRITY (2020–2024)

Variable	USA	Germany	France	Japan	United Kingdom
AI Adoption (AI)	0.084*** (0.021)	0.062*** (0.018)	0.058*** (0.017)	0.041** (0.019)	0.071*** (0.020)
Explainable AI (XAI)	0.126*** (0.028)	0.143*** (0.031)	0.118*** (0.029)	0.097** (0.041)	0.134*** (0.030)
Human-Centered Oversight (HCO)	0.091*** (0.024)	0.104*** (0.027)	0.087*** (0.025)	0.072** (0.030)	0.099*** (0.026)
Embedded Compliance (COMP)	0.153*** (0.035)	0.167*** (0.038)	0.149*** (0.036)	0.162*** (0.040)	0.158*** (0.037)
AI × XAI	0.058*** (0.014)	0.064*** (0.016)	0.052*** (0.015)	0.049** (0.020)	0.061*** (0.016)
AI × HCO	0.043** (0.018)	0.047** (0.020)	0.039** (0.019)	0.036* (0.021)	0.045** (0.019)
Controls	Yes	Yes	Yes	Yes	Yes
Country FE / Time FE	Yes / Yes	Yes / Yes	Yes / Yes	Yes / Yes	Yes / Yes
R^2 (within)	0.71	0.74	0.69	0.66	0.72
Observations	25	25	25	25	25

Notes:Robust standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.Dependent variable: Financial Integrity Index (FI)Estimation method: Fixed Effects (country & time), robust standard errorsSource: author's development using data from (Bank for International Settlements [BIS], 2023; European Commission, 2024; Financial Stability Board [FSB], 2023; International Monetary Fund [IMF], 2024; Organisation for Economic Co-operation and Development [OECD], 2023, 2024; Stanford Institute for Human-Centered Artificial Intelligence, 2024; Transparency International, 2024; World Bank, 2024)

Several key patterns emerge. First, the application of AI by itself is only positively but moderately relevant to the financial integrity. Second, the coefficients that are related to explainable AI and embedded compliance are more significant and larger when compared to baseline AI adoption. Third, the terms of interaction are positive and statistically significant in all the countries, which implies that the effectiveness of AI depends on the governance and oversight structure.

C. Period-by-period dynamics (2020–2024)

The 2020 year is marked by new financial uncertainty and disruption of operations. Within this timeframe, the effects of AI Adoption on financial integrity are estimated to be positive but in a comparatively low level in all countries. The findings suggest that the implementation of AI at an early stage had a supporting role in enabling institutions to handle the complexity of operations instead of providing immediate benefits of integrity.

Conversely, the relationship between Embedded Compliance (COMP) and financial integrity is quite high and significant in 2020, especially in Germany, Japan, and the United Kingdom. This implies that compliance-by-design architectures that had existed before proved very important in absorbing systemic shocks.

The coefficients related to Explainable AI (XAI) are more accentuated in 2021. Those countries that have stronger regulatory frameworks and AI governance guidelines are improving the indicators of financial integrity most importantly Germany and the United Kingdom.

AI x XAI is statistically significant in every country within this time frame, meaning that transparency and explainability began to transform AI implementation into quantifiable gains of risk reduction. This represents a shift in the use of AI in experiments to the more institutionally entrenched uses.

Human-Centered Oversight (HCO) effect can be observed as being strengthened in 2022. The findings mean that the effectiveness of financial integrity continues to grow with the availability of proficient compliance specialists, inside auditors, and risk officers to interpret and confirm AI outputs.

Japan and France have relatively better marginal effects of HCO during this period and this is an indication of a governance-based concept where AI systems are fed into and heavily screened by human factor. The AI x HCO interaction term makes it true that AI systems yield more integrity products when incorporated in organized oversight designs.

All the key variables and terms of interaction are maximally statistically significant by 2023. This indicates that AI-powered financial integrity engines reach a maturity phase, and technology, compliance architecture and human oversight function as a unified system.

The interaction effects observed between the United States and the United Kingdom are particularly high, as they illustrate the high level of explainable models' adoption and developed internal control infrastructures. The model explanatory power (as part of R^2) also gets stronger in all countries, which means that there is a closer connection between the implementation of AI governance-consciousness and financial integrity.

The magnitudes of coefficients plateau in 2024, which indicates that the marginal returns to increasing AI expansion may not be increasing. Nevertheless, the long-term value of XAI, COMP, and interaction terms proves that the quality of governance is a critical variable.

The profiles of Germany and the United Kingdom are the most balanced as the contribution of AI adoption, explainability, and oversight are rather similar. Japan has high compliance effects and the United States continues to experience gains through the interaction effects and not the independent AI strength.

D. Cross cutting comparison analysis.

Government-based model (United Kingdom, Germany). The impact of Explainable AI and Embedded Compliance is strongest in these countries implying that the effect of AI on financial integrity is most evident when the latter is in line with the formal regulatory and institutional frameworks.

Japan and France have an oversight-based model. In such instances, the Human-Centered Oversight will have a relatively bigger part, which demonstrates the conservative approach toward AI integration based on expert opinion and procedural oversight.

Conditional but technologically-based model (United States). The U.S. has a high intensity in AI adoption, however, it is evident that AI is not enough. Integrity gains come to fruition majorly through interaction effects with explainability and oversight mechanisms.

E. General conclusions from the results

Overall, the empirical findings provide strong evidence that AI-enabled Financial Integrity Engines are effective only when embedded within explainable, compliant, and human-centered governance architectures. AI adoption without transparency or oversight yields limited improvements, while the combination of AI, explainability, and institutional control produces substantial and statistically robust gains in financial integrity.

These results empirically validate the core principles of the IFTF™© methodology, particularly the emphasis on integrity-by-design, human-in-the-loop governance, and compliance as a structural component rather than a post-hoc constraint. The findings also suggest that future financial transformation strategies should prioritize governance-aware AI deployment over purely technological scaling.

V. DISCUSSION

The results of the research evidence an excellent empirical evidence of the hypothesis that governance architecture, instead of individual implementation of artificial intelligence technologies, defines the outcomes of financial integrity. In the economies under analysis, AI uptake has a medium impact on transparency and risk reduction only, in contrast to explainability, embedded compliance, and human-centered oversight which have a strong influence on the integrity results. This finding extends and deepens previous governance-based views of financial and organizational studies focusing on the existence of structured control systems rather than simply on technical optimization.

With this aspect, the research is consistent with Mazur et al. (2023), who show that financial stability and performance, in particular a capital structure management, are not direct functions of the isolated financial instruments, but rational and institutionally embedded models of management. The current results, just like theirs, indicate that AI-based financial systems can create sustainable value only by embedding them into consistent governance systems that harmonize decision rules, accountability and risk controls. Just as the capital structure instruments, AI is a facilitating mechanism whose usefulness lies in its design consistency and not the intensity of its use.

The empirical findings are also consistent with sustainability-based governance studies. Prokopenko et al. (2024) demonstrate that new models, especially those that are green entrepreneurship, can have a significant social and

economic effect, but only when integrated in an institutional structure of coordination. Similarly, the current paper concludes that AI leads to financial integrity in a similar fashion as an independent innovation, but as a subset of a larger institutional ecosystem that involves compliance, transparency, and human controls. This supports the fact that technological innovation and governance maturity are complementary as opposed to substitutive powers in sustainability value generation.

In terms of technological risks, the findings are in strong support of the current developments in explainable AI in fraud detection. As Sodnomdavaa and Lkhagvadorj (2026) show, integrated machine learning-XAI systems are more effective than opaque models in terms of identifying financial statement frauds since they allow for interpretability and auditability. Their premise that explainability is not an instrumental addition to a technical setting, but a structural requirement of financial integrity and regulatory trust is empirically validated by the positive and statistically significant interaction between AI adoption and explainability that was found in this study: explainability is not a technical addition, but a structural precondition of financial integrity and regulatory trust.

On the same note, the results are in line with Rodriguez Valencia et al. (2025), as the systematic review on AI-based compliance in cryptocurrency exchanges identifies explainability and governance as factors determining reduced fraud risks. Nevertheless, their analysis is limited to technology mechanisms when the current study goes further to show that the AI implications of governance are not exclusive to the new markets of digital assets but are also found in the context of the traditional financial systems. This cross-domain consistency enhances the external validity of AI models based on governance.

The relative aspect of the findings also corresponds to Yazdi et al. (2024), who state that the efficiency of the AI-enhanced risk management can differ significantly, based on the institutional environment and the maturity of risk governance. The country-level effects differentiated in the current research, especially the more substantial level of interaction effects in the jurisdictions where more advanced compliance infrastructures were in place, affirm that AI facilitates the management of risks only when followed in the context of well-developed institutional settings. This observation interferes with notions of techno-determinism and promotes a contingency-based perspective of AI usefulness.

Research results are also put in perspective by ethical and trust considerations. Thurzo (2025) proposes the notion of a reliable ethical firewall by putting the emphasis on explaining and controlling as the safeguarding layers against the use of algorithms to exploit them. The ongoing relevance of explainability and human-related monitoring in this research gives empirical support to this conceptual model evidence that the ethical protection becomes quantifiable integrity outcomes as opposed to being normative aspirations.

Similarly, the findings of the study align with the synthesis provided by Gunasekara et al. (2025), who list the principles of transparency, accountability, and human control as the main pillars of responsible AI implementation. This framework is developed further in the current study through the quantification of the interaction between these pillars and AI adoption to achieve high-quality financial integrity results. The findings instead of a compliance checklist, the responsible AI will be a performance-enhancing governance architecture.

This study also has an empirical support of user-centric trust and threat mitigation frameworks. Kafali et al. (2024) contend that user centric and institutional risk point of view should be included in trustworthiness assessment. The benefits of human-centered controls that are found here support their opinion and indicate that AI systems become legitimate and efficient in cases where decision-making is decentralized among algorithms and responsible human actors.

The generalizability of the results can also be supported by insights of the surrounding fields. In his article, Bouderhem (2024) focuses on the context of AI sensors in healthcare and points out that integrity in safety-critical situations relies on explainability and transparency. The similarities to financial systems can be identified: opaque AI in both instances compromises trust and accountability, whereas transparent architectures allow responsible decision-making in the face of uncertainty.

Lastly, the findings are consistent with the infrastructure-level governance solutions, e.g. Rahman et al. (2024), who suggest blockchain-supported AI models to manage risks better. Although they are concerned with technological strengthening of trust, the given study posits that the institutional governance is the deciding layer, whether it is reinforced by blockchain technology, explainable models, or any other technologies. Technology can be an addition to governance and it cannot substitute governance.

Altogether, the discussion shows that there is a definite overlap in different streams of the literature: AI can be the source of financial integrity only in the cases when it is built into the transparent and explainable systems controlled by humans. The current research builds upon the existing knowledge by offering cross-country econometric data that confirms such a postulation and also incorporating the research on finance, sustainability, risk management, ethics, and responsible AI into a cohesive analytical framework. This way, it will contribute to the comprehension of AI-powered Financial Integrity Engines as systems that are governed but not systems that are technologically independent.

VI. CONCLUSIONS

This paper confirms the idea that the effects of financial integrity in the modern financial system are largely explained by the nature of governance architecture, which is not the strength of AI implementation or the sophistication of algorithms. Empirical evaluation of AI-based financial systems that are used within the years 2020-2024 indicates that explicable, compliant and human-centered AI-based settings produce quantifiable and sustainable enhancements of openness, risk management and trust between institutions. The results indicate that AI systems that are governance-conscious

are always more effective than technology-oriented solutions, especially in environments that are characterized by greater levels of uncertainty, regulation, and systemic risk.

The findings show that the adoption of AI in its own right has a moderate effect on financial integrity, but its efficacy is enhanced significantly under the combination of explainable modeling, inbuilt compliance, and designed human control. The best marginal effects are the interactions between the adoption of AI and explainability, which increases the results of integrity, improving auditability, decision transparency, and accountability. Human-based control also supports these impacts by providing interpretation, validation, and contextualization of the outputs of algorithms to the existing financial control systems. This evidence shows that financial integrity is a systemic and institutionally controlled process and not a direct technological product of predictive accuracy or speed of automation.

Theoretically, the work adds to the existing literature on the topics of financial governance, FinTech, and institutional economics since it incorporates explainable artificial intelligence in an analytical system of financial integrity based on governance. The results confirm the thesis that AI can be treated as a conditional governance tool, the effects of which on risk and transparency are facilitated by institutional design. The research takes the concept of AI-enabled Financial Integrity Engines as embedded elements of compliance designs as opposed to autonomous decision-makers, which confines technology-based narratives and redefines AI as an institutional resource whose performance relies on consistency between technology, regulation, and human judgment.

The implications of this study on practice are immense. The findings imply that the implementation of sustainable AI in finance would entail the intentional investment in the explainability standards, compliance-by-design systems, and the human-in-the-loop governance frameworks. The implementation of technology by financial institutions, regulators, and policymakers should be accompanied by transparency and auditability and oversight abilities. When organizations integrate AI into coherent governance systems, they will have a greater chance to minimize their exposure to risks, increase the credibility of their regulatory activities, and preserve the trust of stakeholders. On the other hand, AI applications that do not consider explainability and human accountability will experience diminishing returns and increase compliance risks.

Simultaneously, it is noted that the research has a number of contextual limitations. The financial AI systems are in a fast-changing regulatory, technological, and institutional context. Further studies ought to broaden the longitudinal nature of the analysis, include data on firms and look at newer financial systems where the infrastructures governing them remain in the emergent stage. Additional developments can be made in the dynamic panel modelling, quasi experimental designs and greater exploration of maturity of governance and explainability fidelity. In general, the article finds that strong financial integrity would arise when governance-based AI designs are laid out, where transparency, compliance, and human controls are constructed and reinforced structurally, and not as controls of secondary or post hoc value.

Acknowledgments: None.

Conflicts of Interest: The authors declare no conflict of interest.

Patents: None.

VII. REFERENCES

Aljunaid, S.K.; Almheiri, S.J.; Dawood, H.; Khan, M.A. Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. J. Risk Financial Manag. 2025, 18, 179. https://doi.org/10.3390/jrfm18040179

Bank for International Settlements (BIS). Sound Practices: Implications of Fintech Developments for Banks and Bank Supervisors. BIS Basel 2023. https://www.bis.org/bcbs/publ/d575.htm

Bouderhem, R. A Comprehensive Framework for Transparent and Explainable AI Sensors in Healthcare. Eng. Proc. 2024, 82, 49. https://doi.org/10.3390/ecsa-11-20524

Chen, P.; Wu, L.; Wang, L. AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. Appl. Sci. 2023, 13, 10258. https://doi.org/10.3390/app131810258

Choowan, P.; Daovisan, H. Artificial Intelligence in Data Governance for Financial Decision-Making: A Systematic Review. Big Data Cogn. Comput. 2026, 10, 8. https://doi.org/10.3390/bdcc10010008

European Commission. Digital Economy and Society Index (DESI). European Union 2024. https://digital-strategy.ec.europa.eu/en/policies/desi

Financial Stability Board (FSB). Artificial Intelligence and Machine Learning in Financial Services. FSB 2023. https://www.fsb.org/2023/11/artificial-intelligence-and-machine-learning-in-financial-services/

Gunasekara, L.; El-Haber, N.; Nagpal, S.; Moraliyage, H.; Issadeen, Z.; Manic, M.; De Silva, D. A Systematic Review of Responsible Artificial Intelligence Principles and Practice. Appl. Syst. Innov. 2025, 8, 97. https://doi.org/10.3390/asi8040097

Hohma, E.; Lütge, C. From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. AI 2023, 4, 904-925. https://doi.org/10.3390/ai4040046

International Monetary Fund (IMF). FinTech Notes and Artificial Intelligence in Financial Services Database. IMF 2024. https://www.imf.org/en/Topics/Fintech

Kafali, E.; Preuveneers, D.; Semertzidis, T.; Daras, P. Defending Against AI Threats with a User-Centric Trustworthiness Assessment Framework. Big Data Cogn. Comput. 2024, 8, 142. https://doi.org/10.3390/bdcc8110142

Lastrucci, A.; Pirrera, A.; Lepri, G.; Giansanti, D. Algorethics in Healthcare: Balancing Innovation and Integrity in AI Development. Algorithms 2024, 17, 432. https://doi.org/10.3390/a17100432

Mazur, V.; Koldovskyi, A.; Ryabushka, L.; Yakubovska, N. The Formation of a Rational Model of Management of the Construction Company's Capital Structure. Financial and Credit Activity: Problems of Theory and Practice 2023, 6, 128–144. https://doi.org/10.55643/fcaptp.6.53.2023.4223

Organisation for Economic Co-operation and Development (OECD). Framework for the Classification of Artificial Intelligence Systems. OECD 2023. https://oecd.ai/en/classification

Organisation for Economic Co-operation and Development (OECD). OECD Artificial Intelligence Indicators. OECD 2024. https://oecd.ai/en/ai-indicators

Owens, E.; Sheehan, B.; Mullins, M.; Cunneen, M.; Ressel, J.; Castignani, G. Explainable Artificial Intelligence (XAI) in Insurance. Risks 2022, 10, 230. https://doi.org/10.3390/risks10120230

Prokopenko, O.; Chechel, A.; Koldovskiy, A.; Kldiashvili, M. Innovative Models of Green Entrepreneurship: Social Impact on Sustainable Development

of Local Economies. Economics Ecology Socium 2024, 8, 89–111. https://doi.org/10.61954/2616-7107/2024.8.1-8

Rahman, M.M.; Pokharel, B.P.; Sayeed, S.A.; Bhowmik, S.K.; Kshetri, N.; Eashrak, N. riskAIchain: AI-Driven IT Infrastructure—Blockchain-Backed Approach for Enhanced Risk Management. Risks 2024, 12, 206. https://doi.org/10.3390/risks12120206

Ricciardi Celsi, L. The Dilemma of Rapid AI Advancements: Striking a Balance between Innovation and Regulation by Pursuing Risk-Aware Value Creation. Information 2023, 14, 645. https://doi.org/10.3390/info14120645

Rodríguez Valencia, L.; Ochoa Arellano, M.J.; Gutiérrez Figueroa, S.A.; Mur Nuño, C.; Monsalve Piqueras, B.; Corrales Paredes, A.d.V.; Bemposta Rosende, S.; López López, J.M.; Puertas Sanz, E.; Levi Alfaroviz, A. A Systematic Review of Artificial Intelligence Applied to Compliance: Fraud Detection in Cryptocurrency Transactions. J. Risk Financial Manag. 2025, 18, 612. https://doi.org/10.3390/jrfm18110612

Seralidou, E.; Kioskli, K.; Fotis, T.; Polemi, N. AI_TAF: A Human-Centric Trustworthiness Risk Assessment Framework for AI Systems. Computers 2025, 14, 243. https://doi.org/10.3390/computers14070243

Shaban, O.S.; Omoush, A. AI-Driven Financial Transparency and Corporate Governance: Enhancing Accounting Practices with Evidence from Jordan. Sustainability 2025, 17, 3818. https://doi.org/10.3390/su17093818

Sodnomdavaa, T.; Lkhagvadorj, G. Financial Statement Fraud Detection Through an Integrated Machine Learning and Explainable AI Framework. J. Risk Financial Manag. 2026, 19, 13. https://doi.org/10.3390/jrfm19010013

Stanford Institute for Human-Centered Artificial Intelligence. AI Index Report 2024. Stanford University 2024. https://aiindex.stanford.edu/report/

Thurzo, A. Provable AI Ethics and Explainability in Medical and Educational AI Agents: Trustworthy Ethical Firewall. Electronics 2025, 14, 1294. https://doi.org/10.3390/electronics14071294

Transparency International. Corruption Perceptions Index 2024. Transparency International 2024. https://www.transparency.org/en/cpi/2024

World Bank. World Development Indicators. World Bank DataBank 2024. https://databank.worldbank.org/source/world-development-indicators

World Bank. Worldwide Governance Indicators (WGI). World Bank Group 2024. https://www.worldbank.org/en/publication/worldwide-governance-indicators

Yaseen, H.; Al-Amarneh, A. Adoption of Artificial Intelligence-Driven Fraud Detection in Banking: The Role of Trust, Transparency, and Fairness Perception in Financial Institutions in the United Arab Emirates and Qatar. J. Risk Financial Manag. 2025, 18, 217. https://doi.org/10.3390/jrfm18040217

Yazdi, M.; Zarei, E.; Adumene, S.; Beheshti, A. Navigating the Power of Artificial Intelligence in Risk Management: A Comparative Analysis. Safety 2024, 10, 42. https://doi.org/10.3390/safety10020042