Proceedings of Business and Economic Studies, 2025, Volume 8, Issue 8
http://ojs.bbwpublisher.com/index.php/PBES
Online ISSN: 2209-265X
Print ISSN: 2209-2641

Ethical Risks of Artificial Intelligence in Financial Management: Identification and Governance Based on Stakeholder Theory

Yirun Mao*

Nanjing University of Science and Technology Zijin College, Nanjing 210023, China

**Author to whom correspondence should be addressed.*

Abstract: The in-depth application of artificial intelligence (AI) in the field of financial management (such as intelligent credit scoring and risk control) has significantly improved operational efficiency, but has also highlighted ethical risks such as algorithmic bias and data privacy breaches. Based on stakeholder theory, this paper takes banks and Internet financial enterprises as research objects to systematically identify the manifestations and formation mechanisms of AI ethical risks, and constructs a "technology-institution-ethics" trinity governance framework. The study finds that AI ethical risks are essentially the result of an imbalance in the interests of stakeholders (financial institutions, users, regulators, and technology providers). Algorithmic bias stems from historical discrimination in training data and the "black box" nature of algorithms, while privacy breaches are related to deficiencies in data governance and regulatory lag. Practices such as Microsoft Azure's ethical assessment matrix and the European Union's AI Act demonstrate that the synergy of technological prevention and control, institutional constraints, and ethical consensus can effectively mitigate risks. This paper provides theoretical support for the ethical governance of financial AI and offers references for corporate compliance practices.

Keywords: Artificial Intelligence ethics; Financial management; Stakeholder theory

Online publication: December 31, 2025

1. Introduction

With the widespread application of artificial intelligence (AI) in the financial sector, innovative practices such as JPMorgan Chase's COIN platform and Ant Group's Zhima Credit have become increasingly prevalent. However, ethical incidents have also occurred frequently alongside these developments, such as the credit score discrimination controversy at Wells Fargo and the data breach issue at HSBC, highlighting the urgency of governance in this field. Current academic research predominantly focuses on single types of risks or perspectives from individual stakeholders, lacking a systematic analytical framework based on stakeholder theory, which creates a certain theoretical gap. Meanwhile, the European Union's AI Act has classified financial AI as a "high-risk" area, and regulations such as China's Personal Information Protection Law continue to strengthen data supervision. These regulatory trends collectively point to an urgent practical need: the pressing necessity to construct a governance framework that aligns with the development of financial AI and balances the rights and interests of multiple stakeholders.

2. Literature review

A review of the existing literature indicates that research on the ethical risks of artificial intelligence and their governance is advancing along two parallel yet urgently in need of integration tracks.

In terms of the generation mechanism of ethical risks in AI, research has transcended superficial observations, revealing how technological systems embed and amplify social biases. Taking algorithmic bias as an example, classic studies have identified a transmission chain of "data bias–model amplification–outcome discrimination": historical and structural injustices inherent in initial training data are unconsciously solidified and amplified by machine learning models during the optimization process, ultimately resulting in systematic unfair outputs against specific groups. This process underscores the socially constructed nature of technological risks [1]. Analyses of privacy and data security risks simultaneously focus on technological vulnerabilities (such as model inversion attacks and API interface misuse) and institutional gaps (such as lagging supervision of the data lifecycle), driving the evolution of "privacy by design" from a conceptual framework to practical implementation [2]. Furthermore, the interpretability crisis arising from the "black box" nature of algorithms, the accountability dilemma caused by unclear responsibility chains, and the potential exacerbation of social inequalities through technological applications collectively constitute a multifaceted and intertwined landscape of ethical risks [3].

Meanwhile, stakeholder theory provides a core analytical framework for deconstructing the complex issues involving multiple actors and values described above. Evolving from initial corrections to the "shareholder primacy" paradigm, through legislative practices such as "stakeholder clauses" in U.S. state corporate laws and institutional explorations like the European social enterprise model, the theory has developed into a systemic governance framework emphasizing the identification of diverse values, balancing of interests, and dynamic equilibrium [4]. In the digital age, this theory has been applied to platform and algorithm governance, emphasizing that in the design, deployment, and regulation of AI systems, it is essential to comprehensively consider the rights and responsibilities of diverse stakeholders, including developers, enterprises, users, regulatory bodies, and the broader public [5]. This perspective lays a theoretical foundation for transcending a singular technological or institutional viewpoint and constructing a collaborative governance framework.

However, a comprehensive review of existing research reveals a significant integration gap. Most governance solutions exhibit a path-dependent fragmentation, either focusing on developing technological tools such as privacy-preserving computation and explainable AI for "internal remediation," or relying on external legislation and standards for "end-of-pipe regulation." They have yet to effectively establish a collaborative governance mechanism that deeply integrates "technology-institution-ethics" and runs through the entire lifecycle of the system [6]. Additionally, research cases are predominantly concentrated in specific industries (such as fintech or content recommendation), lacking in-depth cross-industry comparative analyses involving other high-risk domains like financial AI, medical diagnosis, judicial forecasting, and public administration. This deficiency not only hinders a profound understanding of the universal patterns and contextual specificities of AI ethical risks but also restricts the development of more adaptable and resilient governance toolkits.

3. Stakeholder mapping

Tracing the root causes of ethical risks in financial AI back to the dynamic game and structural conflicts among core stakeholders provides a solid theoretical foundation for systematic analysis. Based on the mapping of stakeholder theory, the core actors and their primary demands can be summarized as follows:

(1) Financial institutions, as the dominant and demand-side players, primarily seek to maximize operational efficiency and achieve commercial profit growth through AI;

(2) Users, as recipients of services and providers of data, focus on receiving fair and non-discriminatory treatment and ensuring effective protection of their personal data privacy and security;

(3) Regulatory bodies, as guardians of market order and public interests, aim to maintain the stability, fairness, and transparency of the financial system and prevent systemic risks;

(4) Technology providers (including algorithm developers and platform operators), as key enablers, focus on the rapid deployment of technology, increasing market share, and avoiding relevant legal and ethical responsibilities.

These four entities form a core force network that drives and constrains the development of financial AI. The heterogeneity of their demands and the asymmetry of their power directly give rise to the core conflict areas of ethical risks. There is the "efficiency-fairness" conflict, where financial institutions' pursuit of extreme efficiency through algorithms may come at the expense of user fairness, as exemplified by controversies over interest rate discrimination arising from differential pricing using complex models on platforms like LendingClub. On top of that, it manifests as a game between "innovation and compliance," where financial institutions and even tech companies, in their aggressive pursuit of technological innovation and exploration of business models, often push or even breach existing compliance boundaries.

Ant Group's severe penalties for improper data collection and usage serve as a typical case of such conflicts. The persistence and intensification of these conflicts constitute the underlying logic of risk formation. In the absence of an effective framework for checks and balances and collaborative governance, the demands of powerful stakeholders (such as financial institutions pursuing efficiency) can override the rights of weaker parties (such as individual users) and influence the behavioral choices of intermediaries (such as technology providers). A typical transmission path is that, under the pressure of commercial competition, technology providers, in order to better meet the urgent needs of financial institutions for "cost reduction and efficiency enhancement" and rapid deployment, may simplify or even omit necessary processes for bias detection, interpretability construction, and privacy protection enhancement in algorithm design, testing, and auditing, thereby systematically embedding ethical risks at the technological source.

Therefore, the ethical risks of financial AI are not merely technical glitches; their essence lies in the result of an imbalanced power structure among key stakeholders, unreconciled value objectives, and the resulting distorted behavioral incentives within a specific institutional and market environment.

4. Identification of ethical risks

Based on an empirical analysis of existing typical incidents, ethical risks in the application of financial artificial intelligence can be specifically identified as three interconnected core dimensions: algorithmic bias, data privacy breaches, and their dynamic propagation within the stakeholder network.

Algorithmic bias manifests differently in traditional banking and fintech scenarios. In traditional banking, taking the Wells Fargo case as an example, its credit model used zip codes as a key input feature. Due to the historically racially segregated residential patterns in the United States, this essentially constituted "geographic proxy discrimination" against minority communities, leading to a systematic and unfair allocation of credit resources. In the fintech sector, platforms like LendingClub leverage vast amounts of digital behavioral data (such as browsing duration, social network characteristics, and device models) to construct complex machine learning models, enabling hyper-segmented pricing for users. However, this "personalized" interest rate based on digital footprints often evolves into "algorithmic exploitation" of vulnerable groups due to the high correlation between the data and historical socioeconomic status, in the absence of effective auditing and explanation. The technological roots of this issue can be traced back to the bias inheritance of training data from historical social structures, as well as the implicit reliance on and misuse of "proxy variables" (such as directly linking the "purchase of specific hair styling products" to the probability of default) that are highly correlated with protected attributes like race and gender, in order to enhance predictive accuracy.

The risk of data privacy breaches is exposed to the dual vulnerabilities of technology and systems. Technical vulnerabilities serve as the direct cause; for instance, in the case of HSBC, flaws in its encryption protocols for customer service chatbots resulted in the potential interception of sensitive conversations by third parties, highlighting the security challenges faced by AI systems as new interfaces for data interaction. However, the deeper root lies in the lag and absence of regulatory and compliance frameworks. Taking the Ant Group's penalty case as an example, the crux was its use of user behavior data from non-financial scenarios such as e-commerce and social networking in financial credit evaluation models without obtaining clear, separate, and explicit consent from users. This practice profoundly reveals the fundamental contradiction between the inherent drive of "data assetization" in the fintech business model, maximizing the economic value of user data, and the basic privacy protection principles in legal and ethical norms, such as "informed consent" and "purpose limitation."

Crucially, individual risk points do not exist in isolation but rapidly propagate and amplify across stakeholder networks, triggering a chain reaction of "technical failure–regulatory penalty–market distrust." A complete transmission pathway is clearly visible: Initial algorithmic bias behaviors (such as credit discrimination) trigger investigations and stringent enforcement actions by regulatory bodies (e.g., the U.S. Consumer Financial Protection Bureau imposing a massive $3.7 billion fine on Wells Fargo); following media coverage of the penalty decision, this swiftly transforms into a significant reputational crisis, leading to fluctuations in the institution's stock price and damage to its brand value. Ultimately, this series of events broadly erodes public trust in the fairness and reliability of financial AI systems, thereby elevating the social and compliance costs associated with the industry's development. Empirical analysis reveals that the ethical risks of financial AI represent a systemic challenge arising from the interplay and co-evolution of technological flaws, institutional loopholes, commercial motivations, and social structures, necessitating a thorough understanding through specific case scenarios and dynamic networks of interest game for effective identification.

5. Construction of governance framework

Based on a systematic analysis of the root causes and transmission mechanisms of ethical risks in financial artificial intelligence, this paper advocates for the establishment of a comprehensive "technology–institution–ethics" tripartite governance framework that integrates "hard technological constraints, strong institutional norms, and soft ethical guidance" in a mutually synergistic and dynamically adaptive manner, to systematically address the aforementioned challenges.

5.1. Technical pillar

This pillar focuses on integrating ethical considerations into the system development and deployment processes, primarily driven by technology providers and financial institutions. Its core lies in translating ethical principles into executable technical solutions as follows:

(1) Adopting explainable artificial intelligence (XAI) technologies to enhance the transparency and auditability of algorithmic decision-making processes, transforming "black-box" decisions into understandable and challengeable logical chains;

(2) Promoting privacy-preserving computing technologies such as federated learning and differential privacy to enable collaborative model training and optimization without data leaving its original domain, thereby structurally reducing the risks of privacy breaches and misuse associated with data centralization;

(3) Introducing systematic ethical assessment tools, drawing inspiration from the 12-dimensional assessment matrix developed in Microsoft Azure AI's implementation of ethical principles, to conduct dynamic, lifecycle-wide monitoring and quantitative evaluation of AI systems' fairness, reliability, and security.

5.2. Institutional pillar

This pillar aims to establish a clear and predictable external regulatory and constraint environment, with regulatory bodies and industry associations playing key leading roles. Its construction pathways include as follows:

(1) Implementing risk-based, categorized, and tiered supervision, referencing the EU's AI Act to explicitly classify most financial AI applications as "high-risk" systems and impose stringent compliance obligations accordingly (such as mandatory fundamental rights impact assessments);

(2) Clearly define the chain of legal liability through legislation or departmental regulations. For instance, China's "Measures for the Governance of AI in Finance" have exploratorily delineated the principal responsibilities of various parties, ranging from technology providers to financial institutions, thereby resolving the challenge of liability determination;

(3) Encourage and improve industry self-regulatory mechanisms, support organizations such as the Global Financial Innovation Network in formulating more practical ethical guidelines, and form a regulatory system that combines "hard law" with "soft law."

5.3. Ethical pillar

Focusing on shaping a responsible innovation culture and endogenous constraints requires the in-depth participation and collaboration of all stakeholders. Key measures include as follows:

(1) Implementing comprehensive and tiered ethical training for all personnel, from management to technical developers, internalizing it as professional competence. For example, JPMorgan Chase requires relevant employees to complete a cumulative total of 40 hours of AI ethics courses;

(2) Substantially empowering users through product design to enhance their right to information and control. For instance, PayPal provides users with an AI decision-making control panel, allowing them to view and adjust the core factors influencing their credit scores;

(3) Establishing independent, interdisciplinary, and cross-departmental ethics committees at both institutional

and industry levels, drawing on the pioneering experiences of institutions such as the Bank of Canada, to play a central role in reviewing major AI projects and adjudicating ethical disputes.

In summary, this framework emphasizes that only by achieving the integration of ethics through technological means, setting clear bottom lines through institutional rules, providing continuous guidance through ethical culture, and forming an organic linkage and closed loop among the three can a solid and sustainable foundation be established for the responsible innovation of financial artificial intelligence.

6. Conclusion

This study offers valuable explorations in both theory and practice. In terms of theoretical contributions, this research systematically applies stakeholder theory to the analytical framework for ethical risks in financial artificial intelligence. It not only clearly reveals the inherent mechanisms by which risks emerge from the interactions and game-playing of demands among financial institutions, users, regulatory bodies, and technology providers, but also breaks through the governance perspective of existing literature that mostly focuses on "single risks" such as algorithmic bias or data privacy. Instead, it constructs an analytical paradigm of "systemic governance" based on multi-party interactions and risk transmission, thereby filling the research gap from isolated problem-solving to holistic and interconnected governance.

In terms of practical implications, the research findings provide actionable paths for different stakeholders. For financial institutions, it is essential to shift ethical review from post-hoc remediation to "proactive ethics," deeply integrating requirements such as fairness assessment and explainability design throughout the entire AI system development process. For regulatory bodies, there is a need to establish an agile and dynamic mechanism for updating regulatory rules to keep pace with rapid technological iterations, while also strengthening cross-departmental collaborative regulatory capabilities. For users, it is crucial to actively enhance their digital literacy and AI cognitive abilities to more effectively exercise their rights to information, consent, and objection, thereby safeguarding their legitimate rights and interests.

Certainly, this study also has certain limitations and points to future research directions. On one hand, the case analyses primarily focus on the relatively mature regulatory systems of the US, European, and Chinese markets, and the applicability of the conclusions in emerging economies with vastly different institutional environments (such as Southeast Asia and Africa) requires further examination. On the other hand, there remains room for deeper exploration of technological solutions. Future research could focus on leveraging the immutability and traceability characteristics of blockchain technology, combined with cryptographic schemes such as zero-knowledge proofs, to construct more robust data privacy protection and audit trail mechanisms, thereby providing a more solid technological infrastructure for the "technology–institution–ethics" governance framework.

Funding

General Project of Philosophy and Social Science Research in Jiangsu Higher Education Institutions for the Year 2025 (Project No.: 2025SJYB0344)

Disclosure statement

The authors declare no conflict of interest.

References

[1] Barocas S, Selbst A, 2016, Big Data's Disparate Impact. California Law Review, 104(3): 671–732.

[2] Cavoukian A, 2020, Privacy by Design: The 7 Foundational Principles, viewed May 1, 2024, https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

[3] Freeman R, Harrison J, Wicks A, et al., 2010, Stakeholder Theory: The State of the Art, Cambridge University Press, Cambridge.

[4] Mittelstadt B, Allo P, Taddeo M, et al., 2016, The Ethics of Algorithms: Mapping the Debate. Big Data & Society, 3(2): 1–21.

[5] Zhang X, 2022, Construction of a Stakeholder Collaboration and Responsibility Framework in Algorithm Governance. China Legal Science, 2022(5): 248–268.

[6] Zarsky T, 2015, The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated. Science Technology & Human Values, 41(1).

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.