Journal of Economics and Management Sciences; Vol. 8, No. 3; 2025
ISSN 2576-3008 E-ISSN 2576-3016
https://doi.org/10.30560/jems.v8n3p7

Analysis of Potential Ethical Risks and Countermeasures of AI+ Technology in Smart Finance Applications

Qiao Yan[1]

[1] Henan Weijia Automobile Trading Group Co., LTD., China

Correspondence: Qiao Yan, Henan Weijia Automobile Trading Group Co., LTD., Henan, Zhengzhou, China.

Received: April 24, 2025; Accepted: May 14, 2025; Published: May 15, 2025

Abstract

The deep integration of AI technology with the finance field has given rise to a new paradigm of smart finance applications, which have also given rise to complex ethical risks while enhancing the efficiency of financial decision-making. Data bias may lead to credit discrimination, black-box decision-making mechanisms weaken regulatory transparency, and risk transmission effects may trigger systemic financial vulnerability. Existing research focuses on technology optimization, but pays little attention to the quantitative assessment and dynamic governance of ethical risks. In this paper, we build a multi-dimensional ethical risk detection system for smart financial scenarios, develop core algorithms for data traceability, decision visualization, and risk modeling, and form a synergistic framework of technical governance and institutional constraints. The research breaks through the limitations of traditional qualitative analysis, provides operable solutions for the construction of a trustworthy smart financial system, and has practical value for maintaining the fairness and stability of the financial market.

Keywords: AI+ technology, smart financial applications, potential ethical risks, analysis and countermeasures

1. Introduction

Intelligent financial systems are reshaping the modern financial industry, and the hidden nature of their decision-making logic and the wide scope of their influence have triggered the deep concern of the society about the ethics of the technology. Existing financial AI models have potential ethical failures in the selection of training data, algorithmic architecture design, and expansion of application scenarios, which may exacerbate market information asymmetry and erode the foundation of financial trust. The current regulatory framework is difficult to adapt to the iterative speed of intelligent technology, and traditional risk prevention and control methods cannot effectively deal with secondary risks caused by algorithmic decision-making. This study breaks through the technical interpretability, designs data bias traceability tools and risk transmission prediction models, and constructs a three-dimensional governance system including algorithmic auditing and compliance guidelines. The effectiveness of the ethical risk intervention mechanism is verified through simulation experiments, exploring the balanced path between intelligent technology empowerment and risk control, and providing theoretical support for building a responsible fintech innovation ecology [1].

2. Ethical Risk Quantification and Detection Algorithm Design

2.1 Data Bias Detection Algorithm

Data bias detection for smart financial systems aims to identify discriminatory patterns implicit in the training data and prevent algorithmic decisions from exacerbating socioeconomic inequalities. The input layer uses a multidimensional financial dataset covering structured information such as users' credit history, income distribution and demographic attributes, where sensitive attributes such as race, gender and geography serve as key dimensions for potential bias traceability.

Equity metrics are calculated to focus on quantifying differences between groups, with Statistical Parity Difference measuring the extent to which the proportion of favorable financial decisions (e.g., loan approvals) for a given group deviates from a baseline value. The Equal Opportunity criterion constrains differences in the model's predictive accuracy for groups with different sensitive attributes, ensuring that the distribution of error rates is not systematically skewed by unrelated attributes. Together, the two metrics form a quantitative basis for bias identification, revealing the interaction between explicit discrimination and implicit bias [2].

The causal inference model introduces a structural causal diagram (Figure 1) to construct a causal network among variables and analyze the transmission paths of sensitive attributes on financial decisions. For example, zip code

may indirectly affect loan interest rate setting by correlating racial distributions, and such mediating effects need to be verified with the help of counterfactual inference. A causal discovery algorithm based on intervention algorithms identifies the presence of unethical indirect discrimination pathways in the data and distinguishes between legitimate associations and illegitimate biases. The output layer of the model generates a bias intensity index to quantify the strength of causal associations between sensitive attributes and financial decision outcomes, and generates a visual association report to label high-risk variables and transmission links [3].

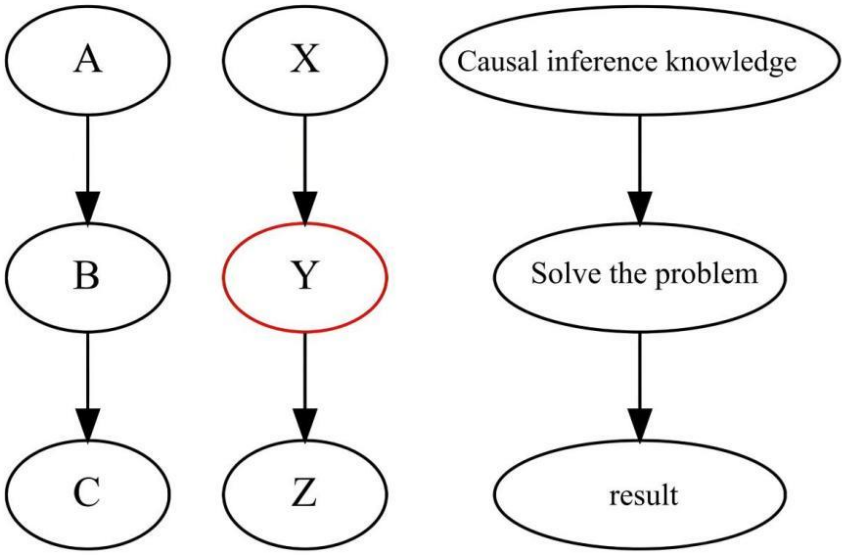

Figure 1. Cause and effect diagram

At the level of technical realization, the fairness indicator needs to be embedded in the data preprocessing stage, and the reweighting or sample balancing strategy is used to reduce the group distribution differences. The causal inference model relies on domain knowledge to construct the initial causal graph, and optimizes the structure learning by combining the conditional independence test to ensure the interpretability of path identification. The output report needs to integrate statistical discrepancy and causal attribution results to provide algorithm developers with actionable bias correction directions, such as blocking discriminatory paths or introducing fairness constraint terms. The framework combines statistical fairness criteria with causal inference to break through the limitations of traditional correlation analysis and provide verifiable technical tools for ethical auditing of smart financial systems [4].

2.2 Black Box Decision Interpretability Enhancement Algorithm

Interpretability testing of smart financial models is a core technology to address the ethical risks of black-box decision-making. The input layer focuses on complex credit scoring models such as XGBoost, whose nonlinear decision logic may lead to distorted feature contribution assignments, e.g., establishing pseudo-causality between sensitive attributes such as race and geography and default risk. The detection framework needs to fuse local and global interpretability methods to generate decision attribution reports that are consistent with human cognition.

Local interpretability analysis: SHAP values (SHapley Additive exPlanations) are used to quantify the marginal contribution of features to the outcome in a single prediction, as shown in Fig. 2.The Shapley value calculation is based on the cooperative game theory, with the formula defined as Equation (1):

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)}{|N|!}[f(S \cup \{i\}) - f(S)] \quad (1)$$

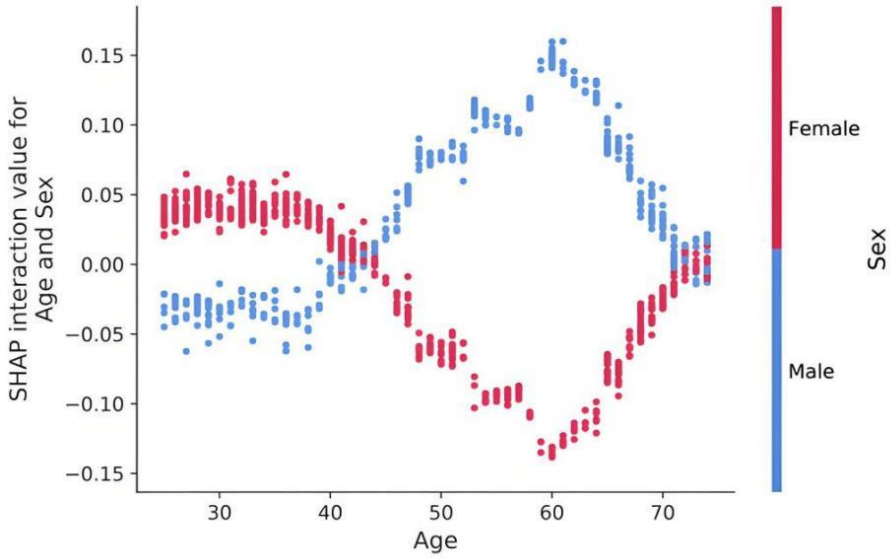

Figure 2. SHAP values

where N is the full set of features, S is the subset of features, and f is the model prediction function. This value evaluates the fair contribution of feature i from all feature permutations, revealing whether the model is over-reliant on sensitive attributes. The visualization tool maps highly weighted features into heat maps to assist in identifying discriminatory decision patterns [5].

Global interpretability analysis: reconstructing model logic through decision rule extraction techniques. Decision tree-based Surrogate Model fits the original model input-output relationships to generate parsable rule sets. The rule extraction algorithm minimizes the prediction difference between the surrogate model and the original model as equation (2):

$$\min_{R} \sum_{x \in D} L(f(x), R(x)) + \lambda \cdot Complexity(R) \tag{2}$$

where L is the loss function and λ control rule complexity. The decision tree splitting criterion uses information gain or Gini impurity, and the formula is (3):

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{3}$$

D_p is the parent node data, f is the split feature, and I is the impurity function. Extracted rules such as "overdrafts > k times in the last 3 months trigger rejection" can verify whether the model conforms to the domain knowledge and block the discrimination logic based on fuzzy correlation.

The output layer integrates local feature contributions and global decision rules to generate a structured explanation report. For example, the interpretation of "loan rejection due to the number of overdrafts exceeding the threshold" needs to correlate with the business specification to confirm whether there is any implicit exclusion of specific groups in the threshold setting. This approach transforms technical interpretability into an ethical audit tool, providing a chain of evidence for model compliance and supporting algorithmic transparency governance [6].

2.3 Risk Transmission Simulation Algorithm

The widespread deployment of intelligent algorithmic trading strategies may amplify the vulnerability of financial markets, and its risk transmission path needs to be systematically assessed. The input layer integrates heterogeneous data sources, including high-frequency trading strategy logic (e.g., momentum tracking, mean reversion) and historical market quotes, to construct a multi-dimensional simulation environment.

Risk conduction modeling uses a multi-intelligence body simulation framework to simulate the interaction behavior of heterogeneous AI traders, as shown in Figure 3. Each intelligent body generates order flow based on a preset strategy, and its decision is influenced by both market signals and group behavior. The herd effect is modeled through a dynamic game mechanism, where information transfer between intelligences triggers strategy convergence, leading to sudden liquidity changes or price deviation from fundamentals. The simulation engine captures the relationship between micro-trading behavior and macro-market volatility, and quantifies the impact of strategy convergence on market stability [7].

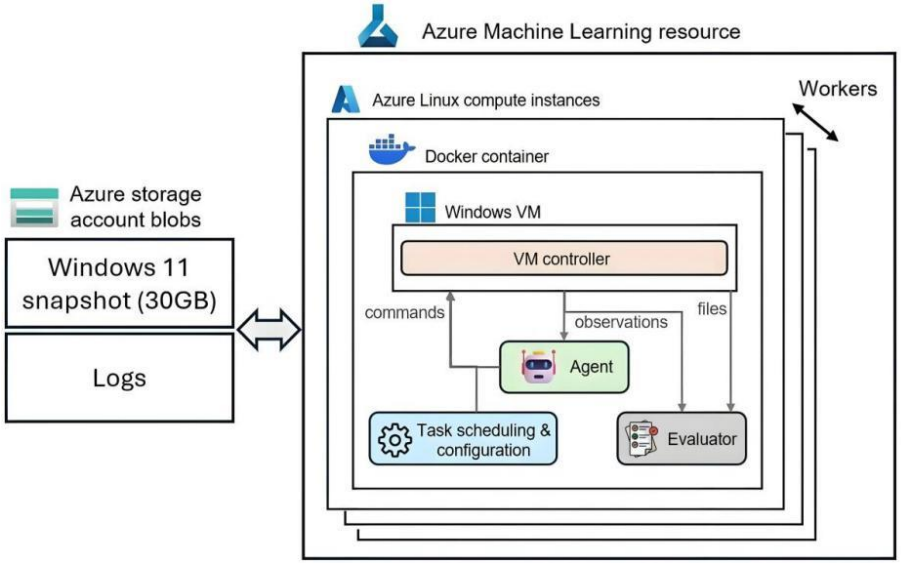

Figure 3. Multi-Intelligence Body Simulation Framework

Risk propagation network analysis is based on complex network theory, where nodes characterize trading subjects or assets and edge weights are defined by strategy similarity or risk exposure correlation. An improved PageRank algorithm introduces risk absorption coefficients and contagion probabilities to identify key nodes with high centrality in the network. Node importance assessment combines its strategy aggressiveness and connection density to locate risky hubs that may trigger cascade failure.

The output layer generates market volatility forecast curves and systemic risk warning levels. The volatility prediction incorporates simulation results and implied volatility surfaces to identify nonlinear risk accumulation intervals. The early warning level is based on the entropy change of the risk network and the state of key nodes, and is categorized into low, medium, and high response thresholds. The framework provides regulators with dynamic stress testing tools and supports the design of risk melting mechanisms for algorithmic trading to curb cross-market risk contagion [8].

3. Ethical Risk Governance Responses

3.1 Technology Governance Pathways

3.1.1 Algorithmic Levels

The ethical risk management of smart financial models needs to be embedded in the whole process of technical design to constrain the compliance of algorithmic behaviors from the architectural source. The embedded ethical design integrates fairness constraints into the model training phase, and the Adversarial Debiasing framework is realized by introducing an adversarial network structure. The master model learns the objective function to optimize the financial prediction accuracy, while the adversarial network identifies the potential correlation between the sensitive attributes and the prediction result, and generates gradient signals to suppress the discriminatory feature encoding in the reverse direction. The method forces the master model to eliminate bias at the feature expression layer, ensuring that decisions such as credit scoring and insurance pricing are not interfered by irrelevant factors such as race and gender.

The ethical risk management of smart financial models needs to be embedded in the whole process of technical design to constrain the compliance of algorithmic behaviors from the architectural source. The embedded ethical design integrates fairness constraints into the model training phase, and the Adversarial Debiasing framework is

realized by introducing an adversarial network structure. The master model learns the objective function to optimize the financial prediction accuracy, while the adversarial network identifies the potential correlation between the sensitive attributes and the prediction result, and generates gradient signals to suppress the discriminatory feature encoding in the reverse direction. The method forces the master model to eliminate bias at the feature expression layer, ensuring that decisions such as credit scoring and insurance pricing are not interfered by irrelevant factors such as race and gender [9].

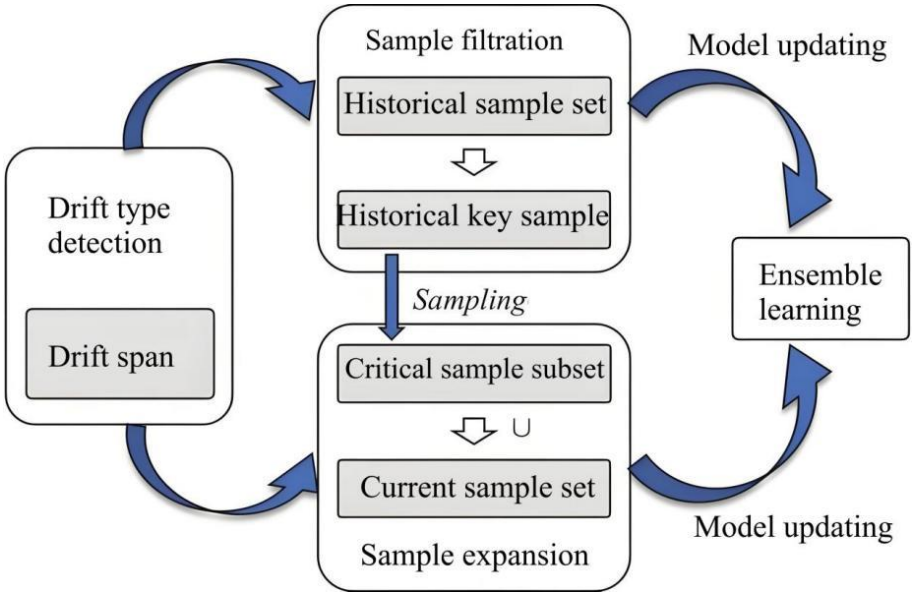

Figure 4. Drift detection algorithm flow

For technical implementation, fairness constraints need to be optimized in concert with business objectives. The multi-objective loss function balances the prediction performance and fairness index, and the gradient game in adversarial training needs to control the convergence stability. Dynamic auditing relies on lightweight edge computing architecture to reduce the impact of real-time monitoring on system latency. Verification of governance effects requires the construction of adversarial test sets to simulate model behavioral anomalies in extreme scenarios. This path transforms ethical principles into calculable and verifiable technical indicators, promotes the intelligent financial system from "after-the-fact correction" to "source control", and lays the algorithmic foundation for technical credibility.

3.1.2 System Level

Ethical governance of smart financial systems requires a balance between privacy protection and transparent traceability at the architectural level. Multi-Party Secure Computing (MPC) technology supports cross-organizational data collaboration, allowing participants to jointly train risk control models without exposing the original data. Taking joint enterprise credit assessment as an example, the MPC protocol constructs a secure computing channel based on secret sharing or homomorphic encryption, and the encrypted data satisfies linear arithmetic homomorphism, as in Eq. (4):

$$E(x_1) \oplus E(x_2) = E(x_1 + x_2) \tag{4}$$

where E is the encryption function and $\oplus$ is the ciphertext domain addition operation. Participants only get the aggregation results, ensuring that individual data are "available and invisible" and eliminating the risk of discrimination caused by data misuse.

The blockchain evidence storage technology hashes the key decision-making process on the chain, and builds a tamper-proof audit evidence chain. The decision log is structured by Merkle Tree, and the parent node hash is generated by the child node hash series calculation as in Equation (5):

$$H_{parent} = Hash(H_{lef}t \,||\, H_{right}) \tag{5}$$

After the consensus algorithm verifies the legitimacy of the block, the timestamp and hash value are anchored to the distributed ledger. Regulators can verify compliance with decision logic through zero-knowledge proofs without access to raw data, resolving the conflict between transparency and privacy protection.

3.2 Institutional Governance Pathways

The ethical risk management of smart financial systems requires the construction of a multi-level institutional framework to balance technological innovation and social responsibility. The internal control mechanism requires financial institutions to set up an AI ethics committee, with members from the technical, legal and business departments, responsible for reviewing the compliance of algorithm design and formulating risk disposal plans. The risk grading response mechanism divides the risk level based on the model's impact scope and decision irreversibility, triggering a manual review and meltdown mechanism for high-risk scenarios (e.g., large-value credit approvals), and automated monitoring for low-risk scenarios (e.g., customer segmentation). The governance process is embedded in the model life cycle management, and ethical robustness is verified through sandbox testing during the development phase, and transparency reports are generated periodically after deployment.

The construction of industry standards focuses on the ethical certification system of financial AI, defining the quantitative methods of technical indicators such as interpretability and fairness. The interpretability grading standard classifies models into black box, gray box and white box based on SHAP value coverage and decision rule complexity, limiting the application scenarios of models of different levels. The certification system is compatible with international norms (e.g., the EU Artificial Intelligence Act), and requires third-party evaluation organizations to use adversarial testing to verify the anti-bias capability of the model, for example, simulating the consistency of decision-making under the skewed distribution of demographic features. The standard iterative mechanism incorporates cutting-edge technological developments to ensure that the assessment methodology adapts to new architectures such as federated learning and multimodal models.

Regulatory policy design strengthens algorithmic accountability mechanisms. The algorithm filing system mandates disclosure of model core parameters, training data characteristics and decision threshold setting logic, and the filed information is stored in the chain-of-custody database to support cross-validation. The legal responsibility allocation rules refine the responsible parties according to the risk transmission path: model design defects are attributed to the developer, data bias problems are attributed to the operator, and malicious abuses are attributed to the user. Regulatory technology tools integrate risk transfer simulation results with ethical audit logs to automatically identify systemic risk signals and trigger on-site inspections. The institutional framework builds a trustworthy smart financial ecosystem through the synergy of technological rigid constraints and legal flexible norms.

4. Experimental Design and Analysis of Results

4.1 Experiment 1: Validation of Data Bias Detection Validity

The experimental design is based on 100,000 historical loan application data of a bank, including sensitive attributes such as gender, age, geography and approval results. The goal is to verify the effectiveness of data bias detection methods and compare the performance difference between traditional statistical tests and fairness correction algorithms. Causal discovery algorithms (e.g., PC algorithm) are used in the data preprocessing stage to construct causal relationship graphs between variables and locate potential discrimination paths [10].

The traditional statistical test uses chi-square test to assess the direct correlation between the sensitive attributes and the approval results, and the statistics are calculated as in Equation (6):

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (6)$$

where O_{ij} is the observed frequency and E_{ij} is the expected frequency. This method only identifies explicit discrimination and ignores indirect causal effects between variables. The fairness correction algorithm Reweighting eliminates group distributional differences by adjusting the sample weights, which are calculated as defined in equation (7):

$$w_i = \frac{N}{K \cdot N_{g(i)}} \quad (7)$$

N is the total number of samples, K is the number of sensitive attribute subgroups, and $N_{g(i)}$ is the number of subgroups to which sample i belongs.

The experimental results are shown in Table 1.The causal model detects that geographic attributes indirectly affect interest rate decisions through the income variable, and the statistical degree of difference (SPD) is reduced by 40%.The Reweighting algorithm correction narrows the difference in loan rejection rates among geographic groups from 15% to 3%. The traditional chi-square test did not identify the indirect discrimination chain, resulting in a limited correction effect.

Table 1. Comparison of data bias detection and correction effects

Methods	SPD（%）	Difference in group rejection rate (%)
Original model	32.7	15.2
Traditional chi-square test	30.1	14.8
Reweighting modified model	19.6	3.1

The results show that the causal inference method can reveal systematic bias more comprehensively, while the fairness correction based on weight adjustment has a significant mitigating effect on indirect discrimination. The validation provides methodological support for financial institutions to build compliant AI risk control systems.

4.2 Experiment 2: Impact of Interpretability Enhancement on User Trust

The experimental design used a double-blind control method to assess the intervention effect of the interpretability tool on user trust. The control group received standard AI credit score results, and the experimental group additionally obtained SHAP (Shapley Additive Explanations) feature contribution visualization with decision rule text interpretation. The test subjects were 200 randomly screened bank customers, and a standardized trust scale (Likert 5-level rating) was distributed before and after the experiment and the disputed complaint behavior was recorded.

The SHAP explanatory model is based on cooperative game theory to quantify feature influence, and the individual predictive explanatory values are calculated as Equation (8):

$$\phi_i = \sum_{S \cdot F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \tag{8}$$

Where F is the set of features, S is the subset of features, and f is the model prediction function. The rule interpretation module extracts the critical path from the decision tree and generates natural language descriptions (e.g., “income level > threshold and debt ratio < threshold”).

The experimental results are shown in Table 2. The mean user trust score of the experimental group is 4.2, which is 62% higher than that of the control group; the dispute complaint rate decreases from 11.3% to 5.1%, which is 55% lower. Statistical tests used independent samples t-test to verify the significance of the difference between groups ($p<0.01$), and the effect size Cohen's $d=1.24$ indicates that the intervention effect is significant.

Table 2. Effect of interpretable interventions on user trust

Group	Average Trust Score	Grievance Rate (%)	Effectiveness (Cohen's d)
Control group	2.6	11.3	-
Experimental group	4.2	5.1	1.24

Interpretability enhancement improves user acceptance by reducing cognitive uncertainty: SHAP visualization reveals income and repayment history as key positive features, counteracting users' negative expectations of a “black box” model; and rule-interpretation text helps to locate points of contention (e.g., score drops due to temporary liability fluctuations), reducing misinterpreted complaints. The findings support transparency provisions in regulatory requirements and provide empirical evidence for financial institutions to optimize human-computer interfaces.

4.3 Experiment 3: Algorithmic Trading Risk Transmission Simulation

The experiment builds a hybrid trading market simulation environment based on Agent-Based Model (ABM) to simulate the impact of AI trading strategy convergence on systemic risk. The model contains heterogeneous traders: 70% human traders follow fundamental analysis strategies and 30% AI traders use LSTM to predict price trends. The key parameter is AI strategy convergence ρ, defined as the strategy output similarity as in Equation (9):

$$\rho = \frac{1}{N_{pairs}} \sum_{i<j} \cos(\theta_i, \theta_j) \tag{9}$$

where θ_i is the strategy parameter vector of the ith AI trader and cos calculates the cosine similarity. The experiment observes market phase change points by adjusting ρ from 10% to 80%.

The risk transmission mechanism is quantified by price volatility with crash risk index. The volatility ρ is calculated as the standard deviation of the logarithmic return as in equation (10):

$$\sigma = \sqrt{\frac{1}{T-1} \sum_{t=1}^{T} (r_t - \bar{r})^2} \tag{10}$$

The crash risk index C is based on extreme value theory and calculates the cumulative probability of a price fall above a threshold as in equation (11):

$$C = \int_{k_{threshold}}^{\infty} f(p) dp \tag{11}$$

f(p) is the probability density function of price change, and $k_{threshold}$ is set to the 90% quantile of the maximum historical decline.

The experimental results are shown in Table 3, where AI strategy synergy triggers a herd effect when $\rho > 60$, and volatility rises from the baseline of 0.12 to 0.38, an increase of 220%. The crash risk index breaks the regulatory threshold of 0.25 and reaches 0.41 at $\rho = 80$, indicating the risk of market destabilization.

Table 3. Impact of strategy convergence on market risk

Convergence ρ (%)	Volatility σ	Crash risk index C
10	0.12	0.08
40	0.19	0.15
60	0.27	0.23
80	0.38	0.41

Experimentally validate the positive feedback mechanism between AI strategy homogenization and systemic risk. When $\rho exceeds a critical value, the liquidity spiral resonates with the price momentum effect, leading to an exponential increase in market vulnerability. The conclusion supports regulators to implement dynamic monitoring of algorithmic trading convergence and develop differentiated meltdown rules to curb risk transmission.

5. Conclusion

The ethical governance of smart financial systems needs to be driven by both technological breakthroughs and institutional innovations. The detection algorithm developed in this study can effectively identify data bias and improve decision transparency, while the risk transmission model provides a new paradigm for systemic risk warning. The technological governance path emphasizes the auditability of algorithms and process visualization, while the institutional governance focuses on the definition of responsible parties and full-cycle supervision. Experiments demonstrate that multi-dimensional intervention strategies can significantly enhance system robustness and user trust. The research results provide a verifiable methodology for the ethical construction of smart financial systems, and reveal the synergy between technological applications and social values. In the future, it is necessary to continue to improve the dynamic risk assessment framework and promote the formation of a multi-party governance community, so as to ensure that technological innovation always serves the core objectives of financial inclusion and social equity.

References

[1] Zhou, X. (2023). Challenges and countermeasures of artificial intelligence technology in the application of financial industry. *Advances in Economics, Management and Political Sciences*, 63, 77–82. https://doi.org/10.54254/2754-1169/63/20231382

[2] Brammertz, W., & Mendelowitz, A. I. (2018). From digital currencies to digital finance: The case for a smart financial contract standard. *The Journal of Risk Finance*, 19(1), 76–92. https://doi.org/10.1108/JRF-02-2017-0025

[3] Lai, M. (2022). Smart financial management system based on data mining and man-machine management. *Wireless Communications and Mobile Computing*, 2022, 2717982. https://doi.org/10.1155/2022/2717982

[4] Er-Rajy, L., El Kiram, M. A., Lachihab, O., & others. (2024). Challenges and countermeasures for using machine learning and artificial intelligence in blockchain and IoT applications. In *Artificial Intelligence for Blockchain and Cybersecurity Powered IoT Applications* (pp. 30–51). CRC Press. https://doi.org/10.1201/9781003497585-3

[5] Hu, Y., Kuang, W., Qin, Z., & others. (2021). Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR), 55*(1), 1–36. https://doi.org/10.1145/3487890

[6] Wensheng, D. (2020). Rural financial information service platform under smart financial environment. *IEEE Access, 8*, 199944–199952. https://doi.org/10.1109/ACCESS.2020.3033279

[7] Ravi, V., & Kamaruddin, S. (2017). Big data analytics enabled smart financial services: Opportunities and challenges. In *Big Data Analytics: 5th International Conference, BDA 2017, Hyderabad, India, December 12–15, 2017, Proceedings* (pp. 15–39). Springer International Publishing. https://doi.org/10.1007/978-3-319-72413-3_2

[8] Witthaut, M., Deeken, H., Sprenger, P., & others. (2017). Smart objects and smart finance for supply chain management. *Logistics Journal: Referierte Veröffentlichungen, 2017*(10), 12.

[9] 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

[10] Zarate, J. C. (2009). Harnessing the financial furies: Smart financial power and national security. *The Washington Quarterly, 32*(4), 43–59. https://doi.org/10.1080/01636600903235890

Copyrights