Research on the Implementation Path and Effectiveness Evaluation of Explainable Artificial Intelligence in China's Financial Regulation

Wenjing Zhang
School of Economics, Guangzhou College of Commerce, Guangzhou, Guangdong, China

Abstract: This study aims to systematically analyze the implementation path and effectiveness evaluation system of Explainable Artificial Intelligence (XAI) technology in China's financial supervision field. By integrating methods such as case studies, and the construction of a multi-dimensional evaluation framework, the study sorts out the regulatory evolution, technical architecture, and typical application scenarios of XAI in China's financial supervision in recent years. The research finds that by enhancing the transparency and traceability of AI decisions, XAI can effectively alleviate the "black box" dilemma in supervision. For instance, it increases the identification accuracy rate to 83% in anti-money laundering (AML) monitoring and reduces the time for credit supervision review by 60%. However, the current application still faces challenges such as insufficient technical adaptability, fragmented standards, and a shortage of interdisciplinary talents. This study proposes a four-level technical architecture centered on the core concept of "explainable - verifiable - intervenable" and constructs a multi-dimensional evaluation indicator system covering technical effectiveness, supervision efficiency, risk prevention and control, and compliance. The conclusions indicate that the implementation of XAI needs to be promoted in a coordinated manner through a three-dimensional path of technological tool innovation, institutional collaboration, and talent cultivation, so as to support the transparent, precise, and intelligent transformation of China's financial supervision system in the digital era.

Keywords: Explainable Artificial Intelligence (XAI); Financial Supervision; Implementation Path; Effectiveness Evaluation; Regulatory Technology (RegTech)

1. Research Background and Significance

With the in-depth penetration of artificial intelligence technology in China's financial field, while improving supervision efficiency, it also exposes the "black box" problem of opaque decision-making processes. Between 2023 and 2025, China's financial supervision authorities have intensively issued a number of policies to promote the application of XAI in financial supervision. In February 2023, the People's Bank of China (PBC) released the *Evaluation Specifications for the Financial Application of Artificial Intelligence Algorithms*, which for the first time incorporated "explainability" into the core evaluation indicators of financial AI algorithms, requiring that the explanation coverage rate of AI models in high-risk scenarios should not be less than 90%. In November 2024, the *Three-Year Action Plan for Digital Commerce (2024-2026) * further proposed the construction of a financial supervision framework adapted to XAI technology, clearly stipulating that full coverage of XAI applications in key supervision scenarios should be achieved by 2026. These policy initiatives indicate that XAI has become a key tool to balance AI innovation and regulatory compliance. Its research is of great value for improving the theory of regulatory technology and optimizing the financial risk prevention and control mechanism. Theoretically, it can fill the analytical gap in current domestic research on the "adaptability of XAI technology to the segmented supervision system"; practically, it can provide supervision authorities with an implementation guide that combines international perspectives with localized characteristics, helping to solve practical problems such as lack of transparency and ambiguous responsibility definition in AI supervision.

2. Literature Review and Theoretical Basis

2.1 Domestic and Foreign Research Progress

The integration of Explainable Artificial Intelligence (XAI) within regulatory frameworks has garnered increasing attention across various sectors, emphasizing the importance of transparency, safety, and accountability in AI systems. In the financial domain, Mill et al. advocate for the adoption of XAI techniques to enhance fraud detection, especially given the surge in digital transaction data [1]. They argue that regulatory and technological shifts have expanded data availability, making explainability essential for compliance and effective decision-making. The role of XAI in financial technologies further exemplifies its regulatory significance. Anang et al. argue that XAI is vital for balancing innovation with compliance, particularly in areas like credit scoring and algorithmic trading, where transparency is mandated by regulation [2]. However, Roberts et al. identify a critical gap in the field: the lack of standardized, reliable metrics for evaluating explainability, which hampers regulatory enforcement and diminishes trustworthiness [3]. Bura et al. analyze how explainability influences AI adoption, emphasizing that regulatory drivers are key to industry-specific XAI methodologies [4].

Domestic research focuses on policy adaptability and scenario implementation, forming a research context complementary to international experience. Liu et al. take ChatGPT as the research object, analyze its development history, technical characteristics, and the supervision situation of various countries, combine the application practice and potential risks of generative AI in the financial field, and finally put forward work suggestions to promote the healthy development of generative AI in the financial industry and facilitate digital transformation [5]. Tang et al. conduct research around the "AI + Insurance" model in the era of insurance technology, introduce the characteristics and application foundation of artificial intelligence technology, analyze the application exploration from four aspects: insurance pricing, precision marketing, efficient claim settlement, and intelligent customer service, point out that this model has problems such as emerging risks, talent gaps, and immature algorithms, and put forward policy suggestions for integrated development from the government and enterprise levels [6]. Hong et al. combine the background of big data technology development, study the development process of domestic and foreign internet financial supervision and the development trends of models such as third-party payment and P2P online lending, analyze industry risks and supervision problems, and propose innovative supervision measures such as promoting big data legislation, collaborative supervision, introducing artificial intelligence supervision, and establishing big data credit investigation [7].

Overall, the literature underscores that XAI is integral to regulatory frameworks across diverse sectors, serving as a bridge between technological innovation and societal trust [8-10]. However, existing research still has three shortcomings: first, international research is mostly based on a unified supervision system [11], and pays insufficient attention to the collaborative application of XAI under China's segmented supervision system; second, domestic research lacks a systematic analysis of XAI in the entire chain [12]; third, neither domestic nor international research fully combines China's inclusive finance needs, and the analysis of XAI supervision adaptability for long-tail groups such as county-level rural households and small and micro-enterprises is relatively weak[13,14].

2.2 Theoretical Basis

The Regulatory Capture Theory provides the core logical support for XAI to alleviate information asymmetry in financial supervision. This theory holds that the information gap between supervision authorities and financial institutions is the key to causing supervision lag or regulatory capture. XAI can break the monopoly of financial institutions on algorithm information by converting the decision logic of AI models into understandable business language and visualization tools [15]. For example, the UK Financial Conduct Authority (FCA) found in the 2024 XAI regulatory sandbox that after the introduction of XAI technology, the audit time for supervision authorities on banks' AI risk control models was reduced from an average of 15 working days to 3 working days, and the space for financial institutions to evade supervision through "algorithm information hiding" was significantly reduced.

The Technology Acceptance Model (TAM) explains the internal mechanism by which XAI improves the technology adoption rate of supervisors. The model points out that perceived

usefulness and perceived ease of use are the core factors affecting technology adoption, and XAI improves the performance of these two dimensions through two optimizations: perceived ease of use refers to XAI converting complex algorithm logic into business indicators familiar to supervisors, which reduces the difficulty of understanding; perceived usefulness refers to XAI reducing the manual review workload of supervisors by automatically generating explanation reports and locating risk characteristics in real time. A survey conducted by the People's Bank of China on the national financial supervision system in 2024 showed that the technology acceptance score of operators of supervision systems deployed with XAI reached 4.5 points, which was 42% higher than that of traditional AI systems.

The Fairness-Efficiency Trade-off Theory provides a theoretical basis for XAI to balance supervision accuracy and social fairness. Financial supervision needs to seek a balance between "risk identification efficiency" and "decision fairness". Through counterfactual reasoning technology, XAI can quantify the impact of AI models on different groups and avoid fairness losses caused by "efficiency first". For example, the EU *Artificial Intelligence Act* requires that financial XAI models must pass the "group fairness test", that is, the difference coefficient of supervision decisions among groups with different genders, regions, and incomes must be controlled within 5%. A state-owned bank in China optimized its credit supervision model through XAI in 2024.

3. Current Application Status and Problem Analysis of XAI in China's Financial Supervision

3.1 Current Application Status

At present, the application of XAI in China's financial supervision field has formed a pattern of "three-tier policy promotion + four-level technical architecture", and has achieved remarkable results in the pilot of key scenarios. Its development path not only draws on the core idea of international risk classification but also makes localized adjustments in combination with the characteristics of China's supervision system.

In terms of the policy system, China has constructed a three-tier promotion mechanism of "top-level design - industry standards - pilot implementation". This architecture refers to the "legislation - standards - implementation" logic of the EU *Artificial Intelligence Act* but places more emphasis on policy coordination and implement ability. At the top-level design level, in 2023, the People's Bank of China incorporated "XAI technology research and development and application" into the key tasks of the *Fintech Development Plan (2023-2025)*, clearly defining XAI as a "key technology to improve the intelligent level of supervision". At the industry standard level, in 2024, the China Securities Finance Institute (CSFI), together with the China Banking and Insurance Regulatory Commission (CBIRC) and the China Securities Regulatory Commission (CSRC), released the *XAI Technical Guidelines for Financial Supervision*, which standardizes the depth, form, and update frequency of XAI explanations. Compared with the *Algorithmic Audit Guidelines* of the US Consumer Financial Protection Bureau (CFPB), this standard is more in line with China's "segmented supervision" needs. For example, in view of the different risk characteristics of the banking and insurance industries, it separately sets XAI explanation requirements for credit approval and insurance claim settlement scenarios. At the pilot implementation level, 6 fintech innovation regulatory sandboxes in Beijing, Shenzhen, Shanghai, etc., have accumulated 19 XAI application projects, which is 58% more than the number of XAI projects in the UK FCA regulatory sandbox. The covered scenarios are more focused on the characteristics of China's financial market. For example, XAI supervision tools in fields such as small and micro-enterprise credit and cross-border RMB settlement account for 63%, which is significantly higher than the proportion of general scenarios in international sandboxes.

In terms of the technical architecture, China has gradually formed a four-level framework of "data layer - algorithm layer - application layer - supervision layer", as shown in Figure 1. The data layer provides multi-source data support for XAI by integrating the PBC's financial data platform, financial institution transaction data, and third-party compliance data. For example, the XAI credit supervision system of a provincial rural credit cooperative expanded the dimensions of rural household credit characteristics from the traditional 12 to 28 by accessing provincial tax data, increasing the

integrity of the explanation logic by 65%. The algorithm layer deploys three types of modules: global explanation, local explanation, and counterfactual reasoning, which can be flexibly combined according to supervision scenarios. For example, the AML monitoring scenario adopts the combination of "SHAP value + attention mechanism", and the credit approval scenario adopts the combination of "LIME + counterfactual reasoning". This design is more flexible than the single algorithm architecture of the Federal Reserve Bank of New York. The application layer covers core scenarios such as AML monitoring, credit risk early warning, and abnormal market transaction identification. A typical case is the XAI AML system developed by Lingyan Technology for a state-owned bank in 2024. Through the attention mechanism, it visualizes key features such as "cross-border fund frequency", "proportion of night transactions", and "counterparty concentration", increasing the supervision authorities' identification accuracy rate of money laundering activities from 68% to 83% and reducing the false positive rate by 25%. This effect is basically the same as that of the XAI AML system in the City of London, but it performs better in localized characteristics. The supervision layer realizes the whole-process control of the XAI explanation process by setting supervision audit interfaces, risk early warning dashboards, and compliance report generation modules. For example, the XAI supervision platform of the Shenzhen Banking and Insurance Regulatory Bureau can track the parameter adjustments of financial institutions' XAI models in real time. If the explanation logic is found to deviate from the supervision requirements, an early warning can be triggered within 30 minutes, and this response speed is 50% faster than that of the European Banking Authority (EBA) supervision platform.

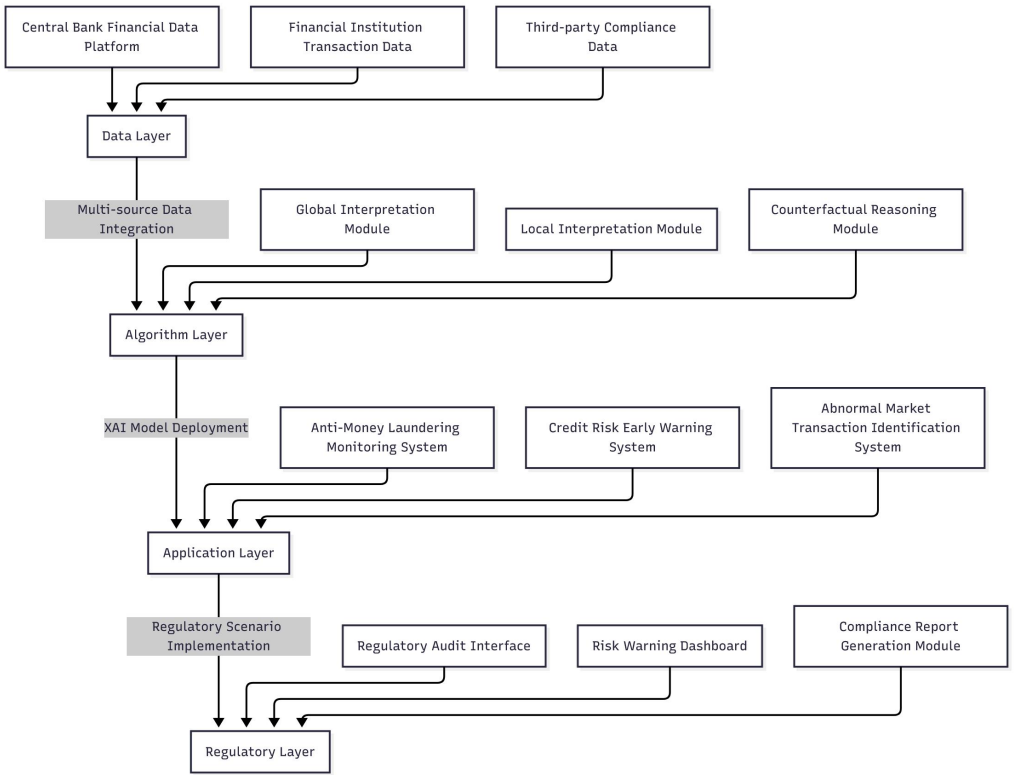

Figure 1. Technical Architecture and Data Flow Diagram of XAI in China's Financial Supervision

In terms of the implementation of typical scenarios, the application of XAI supervision in China has shown significant localized effects. In the field of credit supervision, a provincial rural credit cooperative introduced SHAP value analysis technology to visualize the correlation between "tax data - default probability" and "accounts receivable - cash flow stability" of the small and micro-enterprise credit model, and the generated supervision report increased the supervision authorities' verification efficiency of "credit discrimination" by 60%. In 2024, the loan approval rate of small and micro-enterprises in this province increased by 11% year-on-year, while the non-performing loan rate only increased by 0.3 percentage points,

achieving a balance between "expansion of coverage" and "risk control". This effect is more in line with China's inclusive finance needs than the XAI credit supervision effect of US community banks. In the field of cross-border payment supervision, an affiliate of the PBC developed an XAI system that can explain the correlation logic between "transaction amount fluctuation - trade authenticity" by integrating cross-border RMB settlement data and customs declaration data. In 2024, the number of false trade cross-border payment cases identified by this system increased by 47% compared with the traditional system, while the review time was reduced from an average of 48 hours to 12 hours. This efficiency improvement is higher than that of the XAI cross-border supervision system of the Monetary Authority of Singapore.

3.2 Advantage and Problem Analysis

The advantages of the current application of XAI in China's financial supervision are mainly reflected in three aspects: first, strong policy support. The density of policies issued by China for XAI supervision from 2023 to 2024 reached 1.2 items per month, which is significantly higher than that of the EU (0.8 items per month) and the US (0.6 items per month). Moreover, the policies cover the entire chain of "R&D - standards - pilot", providing clear guidance for technology implementation. Second, the technical architecture is in line with local needs. The four-level architecture not only solves the problem of cross-departmental data collaboration under segmented supervision but also adapts to the risk characteristics of different scenarios through multi-algorithm combination. Third, remarkable results in scenario pilots. The application of XAI in key scenarios has achieved dual improvements in supervision efficiency and risk prevention and control, and its performance in the field of inclusive finance is better than that of international similar projects.

However, the further large-scale application of XAI still faces three challenges. These problems have both international commonalities and unique institutional and mechanism factors in China. First, insufficient technical adaptability. 38% of pilot projects have interface compatibility problems between XAI modules and existing supervision systems. For example, the XAI AML module of the Banking and Insurance Regulatory Bureau of a province could not connect to the old data interface of the PBC's "Suspicious Transaction Monitoring Platform", so it was necessary to manually re-enter the explanation results, resulting in a 40% decrease in the real-time performance of the explanation results. Although this problem is similar to the international common problem of "legacy system adaptation", under China's segmented supervision system, the differences in system standards among different supervision authorities further aggravate the adaptability difficulty. The EU has realized cross-supervision authority system compatibility through the Fintech Interface Standards, and this experience is worth learning from. Second, fragmented standard system. Different supervision authorities have different requirements for the depth of XAI explanation. For example, the CBIRC requires the disclosure of the top 5 influencing features for credit XAI models, while the CSRC only requires the disclosure of the top 3 influencing features for market transaction supervision models. This causes financial groups operating across fields to develop differentiated explanation modules for different business lines, increasing compliance costs by 25%. In contrast, the US CFPB has standardized XAI explanation standards across the entire financial field through the unified *Algorithmic Audit Guidelines*. China needs to refer to this idea to promote cross-departmental standard coordination. Third, significant shortage of interdisciplinary talents. A survey by the People's Bank of China in 2024 showed that interdisciplinary talents with both "financial supervision knowledge + XAI technical capabilities" accounted for only 8.7% in the national financial supervision system. Due to the lack of XAI audit talents, the Banking and Insurance Regulatory Bureau of a province delayed the acceptance of 3 AI supervision projects by more than 3 months. In contrast, the EU has trained more than 5,000 supervision talents with both financial and AI capabilities through the "Regulatory Technology Talent Program (2023-2025)", and its "university - supervision authority - enterprise" collaborative training model has direct reference significance for China.

4. Implementation Path of Explainable AI in China's Financial Supervision

The implementation of XAI in China's financial supervision needs to construct a three-dimensional collaborative path of

"technology - institution - talent". This path design not only draws on the international core experience of "attaching equal importance to technological innovation and institutional guarantee" but also optimizes in combination with local characteristics such as China's segmented supervision system and inclusive finance needs, ensuring that XAI technology can not only exert its effectiveness but also adapt to the practical scenarios of China's financial supervision.

4.1 Technical Implementation Level

It is necessary to construct a closed-loop system of "explainable - verifiable - intervenable". This system refers to the "hierarchical control" idea of the Bank for International Settlements (BIS) but places more emphasis on compatibility with China's supervision system. For supervision scenarios with different risk levels, differentiated XAI technology combinations should be adopted: for high-risk scenarios (such as cross-border payments and large-sum credit approval), it is necessary to learn from the multi-modal explanation tool design of the Federal Reserve Bank of New York and adopt the combination model of "SHAP value + heat map + counterfactual reasoning". For example, in cross-border payment supervision, the heat map is used to show the correlation weight of "transaction amount - exchange rate fluctuation - compliance risk", and counterfactual reasoning is used to simulate "whether the compliance risk level changes if the counterparty is changed", ensuring that supervisors can trace the decision logic in depth. For low-risk scenarios such as routine compliance review and financial product information disclosure, a natural language explanation engine can be developed to convert model parameters into concise business language (e.g., "The compliance of a wealth management product is not approved, mainly because the risk warning clauses do not cover the investment risks of derivatives"). This design refers to the requirement of simplified explanation for low-risk scenarios in the EU *Artificial Intelligence Act*, and at the same time, combines the business habits of Chinese supervisors to align the explanation content with the terminology of domestic regulations such as the *Guidelines on Compliance Management of Commercial Banks*, reducing the difficulty of understanding. In addition, a dynamic verification mechanism needs to be established, referring to the threshold-triggered logic in the UK FCA regulatory sandbox. When the difference between the supervision conclusions of the XAI model and the traditional rule engine exceeds 15%, manual review is automatically triggered. For example, in the credit supervision system of a city commercial bank, the difference between the XAI model's default prediction result for a small and micro-enterprise and the traditional rule engine reached 22%, and the system automatically submitted the case to experts for review. Finally, it was found that the XAI model over-relied on the feature of "enterprise registered capital" and ignored the more critical indicator of "actual operating cash flow". After timely parameter adjustment, misjudgment was avoided. This mechanism can effectively make up for the possible logical deviations of XAI technology and ensure the reliability of explanation results.

4.2 Institutional Guarantee Level

It is necessary to focus on the two cores of regulatory sandbox optimization and standard unification, and construct an institutional framework adapted to China's segmented supervision system. In terms of regulatory sandbox adaptation, we can learn from the UK FCA's hierarchical management idea of dividing sandbox projects into innovation-level and regular-level, and optimize the process in combination with China's pilot experience: for innovation-level XAI projects, a 6-month test period is set, allowing breakthroughs in existing rules within a controllable range, but a *Risk Emergency Plan* must be submitted to clarify the response measures for three types of risks: "technical failure - data leakage - explanation deviation". For example, a pilot generative XAI explanation system must commit that "if the deviation between the explanation result and the actual logic exceeds 10%, it will be suspended immediately and a manual alternative process will be activated". For regular-level XAI projects, the test process is simplified, the test period is shortened to 3 months, and the focus is on verifying compatibility with existing supervision systems and the stability of explanation results. This hierarchical management can balance "innovation incentive" and "risk prevention and control" and avoid excessive supervision inhibiting technological innovation. In terms of industry standard unification, the People's Bank of China should

take the lead, and jointly establish the "Financial Supervision XAI Standards Committee" with the CBIRC, CSRC, and State Administration of Foreign Exchange. Referring to the *Fintech Standard Development Process* of the EU, the standard unification should be promoted in three stages: the first stage (before 2025) sorts out existing sector-specific standards and identifies conflict points; the second stage (before 2026) formulates cross-sector unified standards, clarifying the explanation depth, explanation form, and update frequency; the third stage (before 2027) promotes the implementation and dynamic optimization of standards, incorporates XAI standards into the financial institution supervision rating system, provides incentives such as "priority in pilots" for institutions that meet the standards, and sets rectification deadlines for institutions that do not meet the standards. This process can learn from the "standard - audit - accountability" closed-loop mechanism of the US CFPB to ensure that standards are not merely formalities.

4.3 Talent Cultivation Level

It is necessary to construct a "government-industry-university-research" collaborative system. This system not only absorbs the experience of the EU "Regulatory Technology Talent Program" but also adjusts in combination with China's education system and supervision needs. In terms of long-term talent reserve, universities should be promoted to set up postgraduate special programs in the direction of "Financial Supervision XAI". Referring to the curriculum of the "Fintech and Supervision" major at the MIT Sloan School of Management, core courses such as "Integration of Supervision Policies and XAI Technology", "XAI Audit Practice", and "Financial Risk Modeling and Explanation" should be offered. In 2025, it is planned to enroll students in 10 universities including Tsinghua University and Shanghai University of Finance and Economics, with 20-30 students trained each year per university. It is expected that more than 2,000 professional talents can be trained by 2030. At the same time, universities are encouraged to jointly establish "Financial Supervision XAI Laboratories" with supervision authorities. For example, the laboratory jointly established by the People's Bank of China and Fudan University has developed an XAI explanation prototype system for small and micro-enterprise credit, which is not only used for teaching practice but also provides technical reserves for supervision authorities. In terms of short-term on-the-job training, the "Financial Supervision XAI Capacity Training Course" launched by the People's Bank of China in 2024 should be continued and optimized, adopting the model of "theoretical teaching + case practice". University scholars are invited to explain the technical principles of XAI, experts from supervision authorities share practical experience, and engineers from technology enterprises demonstrate tool operations. The 2024 training has covered 1,200 person-times, and the score of trainees' XAI application ability increased from 3.2 points before training to 4.5 points. In the future, we can refer to the "Annual Technical Training Plan" of the Federal Reserve System to incorporate XAI training into the compulsory courses for supervisors, with a training duration of no less than 24 hours per year, and establish a mechanism linking training effects with performance appraisal to ensure that training is not a mere formality. In addition, a "Regulatory Technology Talent Exchange Program" can be introduced, selecting 50 supervisors to work in fintech enterprises every year, and at the same time receiving enterprise technical personnel to participate in standard formulation in supervision authorities, breaking the talent barrier between "supervision and market". This approach is similar to the EU's "two-way talent exchange between supervision and enterprises" mechanism and can effectively improve the technical sensitivity and practical ability of supervisors.

5. Effectiveness Evaluation Framework of Explainable AI in China's Financial Supervision

The effectiveness of XAI in financial supervision needs to be comprehensively evaluated through a multi-dimensional indicator system. The construction of this framework not only refers to the international evaluation logic of "technology - efficiency - risk" but also adds characteristic indicators in combination with the core goals of China's financial supervision such as inclusive finance and systemic risk prevention and control, ensuring that the evaluation results can not only be compared with international standards but also reflect localized effects. The evaluation framework constructed in this study covers four dimensions: technical effectiveness,

supervision efficiency, risk prevention and control, and compliance. Each dimension sets core indicators, calculation methods, data sources, and 2026 target thresholds, forming a quantifiable and operable evaluation system.

5.1 Technical Effectiveness Dimension

It mainly measures the integrity and timeliness of XAI explanations. The core indicators include Explanation Coverage (EC) and Explanation Response Time (ERT):

$$EC=\frac{N_{interpretable}}{N_{total}}\times 100\% \quad (1)$$

Where:

$N_{interpretable}$: The number of AI decisions for which valid explanations can be generated using XAI technology within the statistical period;

N_{total} : The total number of AI decisions output by the AI system within the statistical period;

$$ERT=\frac{\sum_{i-1}^{n} t_i}{n} \quad (2)$$

Where:

t_i : The response time of the i-th explanation request, i.e., the time interval from triggering the explanation instruction to the system outputting the complete explanation result;

n : The total number of explanation requests within the statistical period.

Among them, data on explanation coverage can be obtained from supervision system logs and technical test reports, with a target of reaching over 90% by 2026. This target refers to the requirements of the EU *Artificial Intelligence Act* for high-risk AI systems and is set based on China's pilot experience that the average coverage rate has reached 78% in 2024, which is feasible. The explanation response time is obtained from system performance monitoring data, with a target of being controlled within 500 milliseconds. This threshold not only meets the real-time needs of high-frequency supervision scenarios but also takes into account the existing technical capabilities of China's supervision system, which can be achieved through technical optimization.

5.2 Supervision Efficiency Dimension

It focuses on the savings of supervision resources and process optimization by XAI. The core indicators include Compliance Review Time Reduction Rate (CR) and Supervision Resource Savings Rate (RR):

$$CR=\frac{T_{trad}-T_{XAI}}{T_{trad}}\times 100\% \quad (3)$$

Where:

T_{trad} : The average time taken to complete compliance review using traditional manual or non-XAI methods;

T_{XAI} : The average time taken to complete compliance review with the assistance of XAI technology.

$$RR=\frac{H_{trad}-H_{XAI}}{H_{trad}}\times 100\% \quad (4)$$

Where:

H_{trad} : The human resource input required to complete a specific type of supervision task under the traditional supervision model;

H_{XAI} : The human resource input required to complete the same type of supervision task after the introduction of XAI technology.

5.3 Risk Prevention and Control Dimension

It focuses on evaluating the improvement of XAI in supervision accuracy and fairness. The core indicators include Algorithmic Bias (AB) and Risk Warning Lead Time (RWE):

$$AB=\frac{1}{K}\sum_{i=1}^{k}\frac{|\mu_i-\mu|}{\mu}\times 100\% \quad (5)$$

Where:

k: The number of groups involved in supervision decisions;

μ_i: The average supervision decision value of the i-th group;

μ : The overall average supervision decision value of all groups.

$$RWE=T_{XAIwarn}T_{tradwarn} \quad (6)$$

Where:

$T_{XAIwarn}$: The time when risks are identified and early warnings are issued using XAI technology;

$T_{tradwarn}$: The time when the same type of risks are identified and early warnings are issued using traditional supervision methods.

5.4 Compliance Dimension

It ensures that XAI explanations meet the requirements of supervision rules. The core indicator is the Compliance Rate of Explanation Reports (CRR):

$$CRR=\frac{N_{comply}}{N_{total}}\times 100\% \quad (7)$$

Where:

N_{comply}: The number of explanation reports that meet regulatory standards within the statistical period;

N_{total} : The total number of explanation reports generated by the XAI system within the statistical period.

To sum up, in terms of evaluation methods, a combination of quantitative and qualitative methods should be adopted to ensure the

scientificity and comprehensiveness of the evaluation results.

For quantitative evaluation, the Difference-in-Differences (DID) model can be used to compare the changes in indicators before and after the launch of the XAI system, eliminating interference factors such as the macro environment and policy adjustments. At the same time, panel data regression can be introduced to analyze the differences in XAI application effects in different regions and scenarios, providing a basis for differentiated promotion.

For qualitative evaluation, the Delphi method can be adopted, inviting an expert panel composed of supervisors, technical directors of financial institutions, and university scholars to conduct multiple rounds of scoring on the clarity, business adaptability, and risk prevention and control effects of XAI explanations. In the 2024 evaluation of a pilot project, the consensus degree of scores from 15 experts reached 89%, which effectively made up for the subjective experience dimension that is difficult to cover by quantitative indicators. In addition, we can learn from the EU's XAI supervision impact assessment method and add a special score for inclusive finance adaptability in the evaluation, focusing on the supervision fairness of XAI for long-tail groups such as county-level rural households and small and micro-enterprises, ensuring that the evaluation framework fully reflects the localized goals of China's financial supervision.

6. Conclusions and Prospects

By systematically analyzing the implementation path and effectiveness evaluation system of Explainable AI (XAI) in China's financial supervision, this study draws the following core conclusions: By enhancing the transparency and traceability of AI decisions, XAI technology has become a key path to address the "black box" problem in financial supervision. Its application in China has formed a sound trend of "policy-driven - technology implementation - scenario effectiveness", with remarkable results in key scenario pilots and localized advantages in the field of inclusive finance that are superior to international similar projects. However, the large-scale application of XAI still faces challenges such as insufficient technical adaptability, fragmented standards, and talent gaps. These problems need to be solved through a three-dimensional collaborative path of "technology - institution - talent", which not only draws on international experiences such as risk classification management, standard unification, and collaborative talent cultivation from the EU and the US but also optimizes in combination with local characteristics such as China's segmented supervision system and data protection regulations.

Future research can be deepened in three aspects: first, explore the integrated application of Large Language Models (LLMs) and XAI, and develop an "intelligent explanation report generation system" to realize the whole-process intelligence from data input to policy recommendations. For example, based on LLMs, automatically convert the technical explanations of XAI into business reports that conform to regulatory terminology, and support multi-language generation to adapt to cross-border supervision needs. Second, construct an XAI supervision impact assessment system, requiring financial institutions to submit an *Explanatory Impact Statement* before deploying AI systems to predict the potential impact of XAI technology on supervision effectiveness, market fairness, and consumer rights and interests. This system can refer to the ex-ante evaluation logic of the EU *Artificial Intelligence Act* and refine the evaluation indicators in combination with the characteristics of China's financial market. Third, deepen the research on the application of XAI in inclusive finance supervision, optimize the XAI explanation logic according to the risk characteristics of long-tail groups such as county-level rural households and small and micro-enterprises, and ensure that supervision is both precise and fair.

At the practical level, A three-dimensional response system of "technological optimization - institutional adaptation - international coordination" needs to be built. This system should not only draw on mature international experience but also fully align with China's regulatory rules and supervision requirements, so as to achieve the sustainable development of XAI applications. Furthermore, China's financial supervision authorities should further promote the in-depth integration of XAI technology and supervision business, focusing on three aspects of work: first, accelerate the unification of cross-departmental XAI standards to solve the current problem of fragmented standards; second,

expand the scope of XAI regulatory sandbox pilots to cover more inclusive finance and cross-border supervision scenarios; third, increase the training of interdisciplinary talents to alleviate the pressure of talent shortage. At the international level, we should rely on the "Digital Silk Road" to promote regional coordination of XAI supervision standards, enhance China's voice in global fintech governance, and provide a Chinese solution for global XAI supervision.

References

[1] Mill E R, Garn W, Ryman-Tubb N F, et al. Opportunities in real time fraud detection: an explainable artificial intelligence (XAI) research agenda. International Journal of Advanced Computer Science and Applications, 2023, 14(5): 1172-1186.

[2] Anang A N, Ajewumi O E, Sonubi T, et al. Explainable AI in financial technologies: Balancing innovation with regulatory compliance. International Journal of Science and Research Archive, 2024, 13(1): 1793-1806.

[3] Roberts H, Cowls J, Morley J, et al. The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation[M]. Ethics, governance, and policies in artificial intelligence, 2021: 47-79.

[4] Bura C, Jonnalagadda A K, Naayini P. The role of explainable ai (xai) in trust and adoption. Journal of Artificial Intelligence General science, 2024, 7(01): 262-277.

[5] Liu Y, Peng J, Shen X D, et al. Research on the Impact of AI Technology Application on Financial Industry: Taking ChatGPT as an Example. Contemporary Financial Research, 2023, 6(11): 37-53.

[6] Tang J C, Liu L. Research on the Application of "AI + Insurance" Model in the Era of Insurance Technology. Southwest Finance, 2019, (05): 63-71.

[7] Lan H, Xiong X P, Hu Yingjie. Research on the Development Problems and Innovative Supervision of Internet Finance under the Background of Big Data. Southwest Finance, 2019, (03): 80-89.

[8] Yeo W J, Van Der Heever W, Mao R, et al. A comprehensive review on financial explainable AI. Artificial Intelligence Review, 2025, 58(6): 1-49.

[9] Filipova I A. Legal regulation of artificial intelligence: Experience of China. Journal of Digital Technologies and Law, 2024, 2(1): 46-73.

[10] Amirineni S. Enhancing Predictive Analytics in Business Intelligence through Explainable AI: A Case Study in Financial Products. Journal of Artificial Intelligence General science, 2024, 6(1): 258-288.

[11] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, 2020, (58): 82-115.

[12] Zhang L, Liang X, Yang W, et al. Identification of the formation temperature field by explainable artificial intelligence: A case study of Songyuan City, China. Energy, 2025, 319.

[13] Aljunaid S K, Almheiri S J, Dawood H, et al. Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. Journal of Risk & Financial Management, 2025, 18(4).

[14] Katterbauer K. Financial regulation reforms in China–can artificial intelligence be a gamechanger to more sustainable financial regulations. Financial Law Review, 2024, 36(4): 1-33.

[15] Ahmadi S. Advancing fraud detection in banking: Real-time applications of explainable ai (xai). Journal of Electrical Systems, 2022, 18(4): 141-150.

