Ethical Dilemmas and Governance Paths of Algorithms in Fintech Innovation

Jinshan Guan*

Shandong University of Science and Technology, Tai'an, 271000, China
**Corresponding author: 19715509057@139.com*

Abstract: *With the increase in the efficiency of services in recent years, many ethical problems have begun to appear in the algorithm system of fintech, such as a lack of transparency, unclear division of responsibilities and price discrimination; prediction inaccuracies have occurred due to sample bias, liquidity suction in high-frequency trading and erosion of privacy boundaries caused by cross-domain data fusion. Based on the analysis of the generation mechanism and transmission logic for the problems mentioned above, this paper puts forward three levels of governance: organising the development process with integrated ethical constraints, an adaptive mechanism for a dynamic interpretability framework, and an algorithm audit and responsibility allocation model based on a reputation system. The above paths form a closed loop from start to finish and offer an organised way to deal with the ethical problems of algorithms.*

Keywords: *Fintech Innovation; Ethical Problems of Algorithms; Attribution of Responsibility; Price Discrimination; Dynamic Interpretability; Reputation Mechanism*

Introduction

Now, algebras are being applied to solve problems in the main areas of Fintech, such as credit risk and payments. Although new technology has increased the efficiency of service, problems in the operation of algorithms have also arisen. The original system of financial ethics is based on the idea of accountability for people in positions of power and information symmetry; however, due to the opacity, dispersion of responsibility and rapid change in algorithmic systems, direct application is not feasible. There are two demands for the study of algorithmic ethical dilemmas and governance paths: First, theoretically, it can clarify the types and origins of these problems, and second, it can fill the research gaps in the field of fintech ethics. In fact, it can provide a reference for the development path of financial institutions and reduce their operating costs. Most of the previous studies have focused on describing individual ethical problems and have not systematically covered the whole process of how these problems occur and are handled. The three types of ethical problems in this paper are a lack of transparency, unclear responsibilities and price discrimination; next, we will discuss how these problems are distributed and aggravated by sample bias, high-frequency trading and cross-domain integration; finally, some solutions to these problems will be put forward, such as reform of the development process, improvement of dynamic interpretability and implementation of market reputation assessment.

1. Theoretical Foundation and Typological Classification of Algorithmic Ethical Dilemmas

1.1 Lack of Transparency in Algorithmic Decision-Making and the Problem of Information Asymmetry

One of the main reasons for the lack of transparency in algorithmic decision-making at fintech companies is that the internal logic of these models is not known to the public. Although deep learning can obtain the decision results of complex models through several layers of non-linear transformations, as we do not know the intermediate results of the computation, there is a serious problem of information asymmetry between financial institutions and end-users. When applying for loans or investment products, users do not know why the algorithm gave a certain credit rating or risk level to them, and they are also not informed of this result. A one-way flow of information does not allow the market to form its own views, and there is also the problem of errors that cannot be corrected after an algorithmic mistake.

The reason there is no openness is that the model is neither accurate nor simple to understand. To increase the accuracy of the prediction, many hidden correlations among the features that cannot be represented in the form of simple linear rules or decision trees need to be taken into account. When the algorithm refuses a credit application, we do not know whether the reason is low income, a bad payment history, bias in the data labels, etc. Another kind of information asymmetry is that the financial institution does not know why there is model drift, such as overfitting or adversarial attacks, because it cannot foresee such changes in the future. Therefore, there will be a reduction in the amount of disclosed information on algorithmic decision-making, and at the same time, the confidentiality of trade secrets and user data must also be guaranteed[1].

1.2 Manifestation of Ambiguity in Responsibility Attribution in Automated Credit Assessment

The first reason for the lack of clarity in allocating the responsibility for errors in automated credit scoring systems is the distribution of accountability among all parties. The previous credit approval process was based on the subjective judgment of loan officers; now, the algorithm system will perform the following few steps: data collection, feature engineering, model inference and threshold setting. If there are any losses resulting from incorrect loan rejections or improper loan approvals, then the data supplier, the model training team, the system operation and maintenance department, and even the original data annotators will be jointly and severally liable. Diffusion of responsibility makes it difficult to determine who is responsible for the harm caused by the algorithm, and it reduces the intrinsic motivation of all people involved to exercise due diligence and verify the results of algorithmic operation.

Another reason for the problem of accountability in automated credit assessment is that the algorithms take a long time to be developed and updated. After the model starts to work, it will continuously collect new transaction data and adjust its parameters; thus, the validation results at the beginning of development may no longer be accurate after three weeks due to changes in the distribution of the data. A poor assessment of a particular loan may be the result of the model's parameter settings six months ago and changes in the current market. Therefore, there is a problem with the chain of responsibility; it cannot be determined whether the loss was due to the original flaws in the algorithm or to failure modes caused by changes in the environment. Therefore, the financial institutions are unwilling to correct the algorithmic errors systematically; generally, the cost of accountability is higher than that of rectification.

1.3 Erosion of Market Fairness by Data-driven Price Discrimination

Based on the behaviour data of users, financial institutions set different prices according to this data for price discrimination; at the core of this is the loss of market fairness caused by obtaining information on how much consumers are willing to pay without their knowledge. Based on all kinds of non-traditional data, such as browsing history and click frequency, device type, charging time, etc., algorithms can construct detailed profiles of people's price sensitivity. For the same financial product or loan, those with a higher risk appetite will be charged more; however, they will be told by the algorithm that it is "risk matching" or "personalised service". The price is not in line with the market, and there will be no extra charges for users.

Another problem of the lack of market fairness is that, based on the behaviour of previous users, algorithms have been practising extended price discrimination. When a user has a short-term cash shortage and has taken out an expensive short-term loan, the algorithm may believe that the user does not have good financial management skills and raise the penalty rates for all other financial products. Recently, there has been a price effect of Matthew; that is to say, groups with less stable finances have borne a higher marginal cost, and the stable premium customers have continued to enjoy a good rate. Based on all the market data, price discrimination will not increase the allocation efficiency of resources; instead, consumer surplus will be transferred to the supply side by exploiting information asymmetry, and thus the inclusiveness and fairness of financial markets will be reduced[2].

2. Generation Mechanisms and Transmission Logic of Ethical Dilemmas in Fintech Algorithms

2.1 Sample Bias in Model Training and the Risk of Inaccurate Financial Forecasting

Sample bias in the model training stage is the problem that the distribution characteristics of the training set do not match those of the target population in actual life. Most of the credit scoring and income prediction models employed by fintech companies are based on historical transaction data from a specific time and place. Generally speaking, the data include people who have already been approved by the old system and do not cover individuals without credit history or those with irregular income. The decision rule will be based on biased data, so there will be systematic prediction errors for underrepresented groups, and thus the financial forecast will be inaccurate. The errors are shown in the estimation bias of default rates, and also in the phenomenon of variance amplification for the same model in different areas or for different groups of customers.

Erosion of the forecast due to sample bias is still a problem. Deploy the model, and then the output results of this will affect the direction of the next data collection. Users who are determined to have a low credit score according to the algorithm will not be offered financial services, and therefore their repayment behaviour data will not be added to the training set of the model. It is a feedback loop, and the bias of the first sample will increase with each model iteration instead of decrease. Another kind of risk to the financial forecast is that it will not be updated in light of changes in variable costs. Give relatively high weight to features that have a large variance in the training data; for example, income from informal channels is strongly related to financial ability but is distributed very uniformly, so it should be downweighted or ignored. The weight logic has failed to achieve the goal of financial forecasting and continues to underestimate the repayment ability of the smallest group of customers.

2.2 Ethical Trade-offs in High-Frequency Trading Algorithms and the Liquidity Suction Effect

To achieve the high execution speed of nanoseconds, the algorithms for high-frequency trading have had to ignore some ethical norms in trading. The typical case is Order Concealment + Liquidity Detection. The algorithm will place a small exploratory order to observe the order book on the other side, quickly execute an arbitrage trade, and then notify the counterparty. Although the current trading rules do not explicitly forbid this type of use of information asymmetry, it will harm the health of the market order book. Due to the small size of the displayed orders, it is not possible to determine whether they are from the actual market or have been artificially generated by algorithms. Another kind of decline in ethics is latency arbitrage; by knowing the order flow of other traders in advance, one can obtain a better execution price earlier and thus profit from information asymmetry[3].

Liquidity suction is a phenomenon in which high-frequency trading algorithms gather and spread liquidity in all parts of the market at different times due to certain reasons in the market. If there is a small price difference for the same asset in different stock exchanges, high-frequency trading can be used to take advantage of this difference and make profits through cross-market arbitrage. Although this behaviour will increase the pricing efficiency of the normal market, it will also result in a vicious circle of liquidity transfer in times of high volatility. After the algorithm finds that the selling pressure in a certain area has increased, it will leave that area in a large number of cases and go to other relatively stable places; therefore, the bid-ask spread at the original location will be wide. Liquidity suction will increase the range of the extent of distortion in price discovery, and thus institutions with slightly slower trading platforms or different risk management standards will have to bear higher transaction costs. As shown in the transmission path above, although the high-frequency trading algorithm can improve the execution quality of individual orders, there is still a risk that liquidity risk will be transferred to market participants who are not as technologically advanced or have a slower response speed.

2.3 Dynamic Erosion of User Privacy Boundaries due to Cross-domain Data Fusion

Cross-domain data fusion is the combination and application of user data that have been collected for different purposes in various places to build a complete profile of that user. With the user's consent, many financial technology platforms have collected transaction data, location history, social graphs, wearable device data, etc. Although the original form of this data does not directly show people's financial behaviour, entity matching and feature cross-referencing can be used to obtain some private financial information about individuals that they have not voluntarily disclosed. Dynamic erosion of the boundary of privacy refers to the expansion, combination and processing of the range of data a user has agreed to provide in one situation, thus more information is disclosed than they were initially aware of when giving consent. Although the users have agreed to the use of their location data by us for anti-fraud verification, they have not been informed that this information will be combined with consumption records to determine how often they see a doctor or whether their marital status has changed[4].

The three stages of the process of privacy boundary erosion are gradual. First, based on the correlation of non-sensitive fields in all the data, attribute inference is used to obtain the sensitive attribute. Secondly, there is the problem of re-identification; if the combined data have many quasi-identifiers, such as timestamps and specific locations, then the original anonymised transaction records can be traced back to particular people. Thirdly, there is a loss of temporal privacy; cross-domain fusion can help us observe changes in people's behaviour over a long period of time, and even combinations of individually inconspicuous actions may indicate serious changes in daily life or financial problems. Given that the erosion environment has changed, we do not know how far the inference ability of the future fusion technology will have advanced by the time of approval. Although the accuracy of prediction by fintech algorithms has been rising with the collection of all kinds of data, the responsibility for protecting privacy has been placed on the end-users; they are generally unaware of or unable to control how their data is used.

3. Governance Paths and Evaluation Framework for Algorithmic Ethical Dilemmas

3.1 Reorganisation Plan for the Algorithm Development Process with Embedded Ethical Constraints

The development plan for the algorithm with ethical constraints is to solve ethical problems in the model design stage, not after the model has been used. The first goal of this paper is to improve the accuracy of prediction and return on investment, but ethical issues were not addressed at that time and have only recently been added through threshold clipping after the model was released. Modify the fairness indicators and demands for openness and traceability of responsibility in the restructuring plan to be measurable constraint functions, and then add them to the loss function in the parameter optimisation stage. The bias-detection module will be added in the data preprocessing stage of the operation stage by the development team to perform an initial check on the distribution of sensitive attributes in the training set and flag any proxy variables that may lead to discriminatory inference in feature engineering. Therefore, the ethics do not need to be added after the fact but should be included in the boundary constraints of the model's parameter space[5].

The other way to construct the restructuring plan is to add a trade-off rule to the multi-objective optimisation problem. Generally speaking, there is a trade-off between prediction accuracy and fairness; that is, increasing the accuracy of credit scoring for one group may lead to an increase in the false-positive rate for another group. The development process should set an upper limit to this trade-off and not strive for the best results in all of the above indicators simultaneously. The Pareto frontier search method can be used in practice to change the model according to different degrees of constraints, and the acceptable deviation threshold will be set by the risk management committee of the bank. A change log of the restructuring plan will also be added to version control to record the reasons for modifying the constraints and the expected results of each round of model iteration. Therefore, the development of algorithms is no longer just a problem of technology but also a problem of engineering and ethics.

3.2 Adaptation Mechanism of the Dynamic Interpretability Framework in Risk Control Models

The framework of dynamic interpretability is not the same as that of static post-hoc explanation; rather, it can adjust the results of the explanation according to the local feature distribution of an input sample. Most of the risk-control models are in the form of complex ensemble structures or neural networks, and the global feature importance ranking cannot explain the reason for a particular credit decision. The local approximation model in the dynamic interpretability framework is used to fit the decision boundary of the original model near the prediction point, and then an explanation vector corresponding to the feature vector of that sample is generated. The adaptation mechanism should place this explanation module in the inference pipeline of the risk control model so that each credit score output is accompanied by a readable attribution analysis report, and this report needs to contain the marginal contribution values of all input features to the scoring result and their respective rankings.

The problems of applying an adaptive mechanism to the risk-control problem are that it is not computationally efficient and lacks good explainability. The generation time of the real-time credit approval decision should be in the order of milliseconds, so the sampling complexity of local approximation methods needs to be small. Therefore, we will pre-calculate some representative anchor points in the feature space and, at the time of online inference, quickly find an approximate model by means of nearest-neighbour search. The framework for dynamic interpretability should also address the problem of explanation instability caused by feature interaction; that is to say, a small change in the input features should not lead to a large fluctuation in the attribution results. A way to make the output of the local explanation model continuous at adjacent samples is to add a smoothness regularization term. This will increase the auditability of the risk-control model and provide specific reasons for the decisions made by the algorithms in different circumstances, rather than just general information about the model.

3.3 Algorithm Audit and Responsibility Assignment Model Based on Market Reputation Mechanism

Another is to use the market's reputation system as an external motivation for algorithm auditing, and at the same time, the behaviour of algorithmic trading by financial institutions can be observed and fed back on by market participants. Traditionally, the audits have been sample-based, and they are carried out at different times by the regulatory department or the company's compliance office. To ensure that the operation of the algorithm is fair, stable and open, a reputation system will be established to publish the results of third-party audits on this algorithm publicly. Both the peer institutions and the end-users will obtain these scores and then decide whether to choose a partner or a financial product. Institutions with a bad reputation will be excluded from the market; thus, capital will be withdrawn and some customers will leave, and this will have a stronger corrective effect on behaviour than internal ethical initiatives.

The Allocation of Responsibility Model determines which party will be held financially responsible for damage caused by algorithms among all the parties in the problem. According to the two dimensions of control rights and beneficiary rights, a calculation formula for the liability coefficient has been added to the model. The extent of alteration that participants can make to the algorithm's parameters and operating environment is called the control rights index, and different weights are given to the model trainer, data provider and deployment institution in this index. The right of the beneficiary will be in line with the amount of economic benefits they have gained from the algorithm. If the wrong credit decision causes losses to the users, the model for distributing responsibility first calculates the product of the liability coefficient and then distributes the compensation amount based on this factor. The model also has a presumption of fault; that is to say, if the algorithm fails the regular audits stipulated in the reputation mechanism, then the burden of proof will be on the financial institution. The two levels of the Design for reputation and quantitative allocation of responsibility make up the internal motivation system for the ethical development of algorithms and are independent of external factors.

Conclusion

Algorithmic ethical problems in fintech innovation are the result of combining the technical characteristics of models with the logic of market resource allocation, and they are not random events. The three main problems are a lack of openness, ambiguous responsibilities and price differentiation; they correspond to the cognitive closure of algorithm systems, the dispersion of participating entities and the concealment of pricing mechanisms. A vicious circle has formed; that is to say, sample bias, the liquidity-suction effect of high-frequency trading, and the erosion of privacy due to cross-domain data fusion are all connected in a chain of risk transmission for the problems mentioned above. Therefore, we will strengthen the ethical constraints and increase the level of governance in the development process, build a dynamic interpretability framework to improve the post-hoc auditability of risk control decisions, and, in line with market reputation, establish an audit-and-responsibility-allocation model and an internal incentive system. In the future, research will be carried out on the dynamic adjustment of the fairness-accuracy trade-off in multi-objective optimisation, various adaptation strategies used by all kinds of financial institutions in light of new ethical constraints, and the development of quantitative privacy risk assessment tools for cross-domain data fusion.

References

[1] Wang Jian, and Qin Xuanzi. What Kind of Financial Services Does Technological Innovation Need:

The Transformation of Financial Paradigm. Southwest Finance, pp.1-11.
[2] Tian Hongzhi, Wang Xiangyu, and Hu Yiyuan. Carbon Trading Regulation, Fintech Empowerment, and Enterprise Green Innovation: Based on the Perspective of Two-Stage Innovation Value Chain. Economic Issues, no.4 (2026):25-34.
[3] Liu Dian, Chen Yunshang, and Shang Xuehan. Research on Financial Support Paths for Deep Integration of Technological Innovation and Industrial Innovation. Credit Reference, vol.44, no.2 (2026):1-11.
[4] Wang Shujuan, Ma Yue, and Li Jiarui. The Green Productivity Enhancement Effect of Sci-Tech Finance: Theoretical Mechanisms and Empirical Evidence. Financial Development Research, no.1 (2026):53-63.
[5] Han Biao. Financial Support for the Integration of Technological Innovation and Industrial Innovation. China Finance, no.24 (2025):19-21.