Big Data Engineering and Intelligent Analysis Framework for Compliance Investigation

Zhen Zhong

Graduate School of Arts & Sciences, Georgetown University, Washington, D.C, 20001, United States
oranosholiday@gmail.com

Abstract: *In the context of increasingly strengthened digital regulation, traditional compliance investigation methods that rely on manual review are no longer able to cope with high-frequency risks and dynamic changes in complex multi-source data environments. This article proposes the construction of a compliance investigation framework that integrates big data engineering and intelligent analysis technology. It systematically outlines the conceptual paradigm and core content of compliance data, designs a data platform architecture with multi-source integration, trustworthy construction, and security management functions, and constructs an intelligent analysis system that integrates behavior recognition, rule generation, and warning decision-making. This study contributes to improving compliance efficiency, enhancing risk prevention and control capabilities, and providing technical support and path references for the digital transformation of organizational compliance governance.*

Keywords: *Compliance Investigation; Data Paradigm; Big Data Engineering; Intelligent Analysis Technology; Dynamic Risk Prevention and Control*

1. Introduction

In the context of high-quality development of the digital economy, data, as a core production factor, widely empowers government, enterprises, and individuals through technologies such as big data, cloud computing, and AI, promoting efficiency improvement and service innovation. However, it also raises complex data security issues. Global data breaches are on the rise, with a 74% year-on-year increase in enterprise data breach cases in 2023. Direct economic losses and trust crises have become key limiting factors. Specifically, at the individual level, the comprehensive analysis of high-value sensitive information such as biometric and trajectory data may form a complete privacy profile, leading to precise fraud and identity abuse; At the enterprise level, the average loss from data security incidents reached 4.88 million US dollars, and small and medium-sized enterprises became a weak link in security due to limited technological capabilities; At the government level, as a dual role regulator and data processor, it faces challenges such as insufficient technological adaptability and difficulties in cross departmental collaboration[1-3].

Traditional compliance investigations rely on manual review, making it difficult to cope with high-frequency risks and dynamic data environments. There are three shortcomings in existing research: personal privacy research often focuses on a single stage and ignores the differentiated characteristics of the entire lifecycle of data (collection, storage, use, sharing, and destruction); Excessive focus on technological protection at the enterprise level, neglecting the hierarchical relationship and impact path of management factors[4-5]; Government policy research often focuses on a single regulation, lacking multi-dimensional systematic evaluations such as policy tools, subjects, and data lifecycle; Multi subject collaborative research and isolated analysis of various parties' behaviors have failed to reveal the strategic interactions and dynamic evolution of the system among the subjects[6-7].

Therefore, this study constructs a compliance investigation framework based on big data engineering and intelligent analysis, aiming to improve compliance efficiency and risk prevention capabilities through systematic methods. The framework design focuses on three dimensions: firstly, based on data cognitive reconstruction, redefining the scope and type boundaries of compliant data, and constructing a data platform architecture that includes multi-source fusion, trustworthy guarantee, and security management; Secondly, an intelligent analysis system that integrates behavior recognition, rule generation, and warning decision-making can achieve cross domain data linkage and dynamic risk assessment; Thirdly, based on the security control requirements of the entire lifecycle of data processing (collection, storage, use, sharing, and destruction), design a multi-level governance mechanism covering technology, management, and policies[8-9].

The main contributions of this study are reflected in three aspects: at the theoretical level, systematically sorting out the compliance data paradigm and core connotations, constructing a data security theoretical system, and forming a complete analytical framework from micro individual mechanisms to organizational management and macro policy governance; At the practical level, establish a privacy concern measurement system covering the entire lifecycle of data, propose a multi-level framework for enhancing enterprise data security capabilities, construct a three-dimensional policy analysis framework of policy tools policy subjects data lifecycle, and create a government enterprise individual tripartite game and warning simulation model; At the methodological level, innovative mixed research methods are employed, combining quantitative analysis (such as structural equation modeling and game simulation) with qualitative research (text coding and multidimensional analysis) to ensure comparability and integration of results from multiple research subjects[10-11].

2. Correlation theory

2.1 The cognitive transformation of data in compliance surveys

Under the framework of compliance surveys, data awareness is undergoing a profound transformation from traditional privacy concepts to modern data security management. Since Warren and Brandeis proposed the groundbreaking definition of "right to solitude" in 1890, the concept of privacy has continuously expanded with technological development. Westin defined privacy as an individual's autonomous control over when, how, and to what extent their own information is communicated to others from the perspective of information control. This definition has been transformed into an emphasis on the ability to control personal data in the information age, gradually focusing privacy protection on the field of personal data security. The international community has formed a broad consensus, with the United Nations Universal Declaration of Human Rights clearly establishing privacy as a fundamental human right, and the European Union's General Data Protection Regulation (GDPR) further defining personal data as "any information about identified or identifiable natural persons", covering both direct and indirect identification possibilities, and providing a comprehensive legal framework for modern privacy protection.Data security and privacy protection are highly integrated in practice: at the technical level, encryption, access control, data anonymization, and other technologies are not only key means of ensuring information system security, but also important tools for protecting personal privacy; At the legal level, GDPR regards data security as a fundamental principle for personal data processing, requiring data controllers to implement appropriate technical organizational measures. Technological innovations such as homomorphic encryption and secure multi-party computation not only ensure data security, but also open up new paths for data sharing and utilization for privacy protection.This cognitive shift is systematically reflected in the theory of data lifecycle. The data lifecycle, as a management framework that describes the complete development process of data from generation to extinction, has the core value of transforming data management from fragmented and passive to systematic and proactive. By planning and controlling the entire data process, it better ensures data quality, enhances data security, and unleashes data value. Although there is diversity in the stage division of data lifecycle in different application scenarios (from three stages to nine stages), they all follow the basic logic of data creation to extinction. This study adopts a five stage model of "collection storage use sharing destruction", which avoids the simplification defects of the three-stage model and the implementation complexity of models with more than six stages, achieving a balance between theoretical completeness and practicality. This model has clear semantics and distinct stage boundaries, and is suitable for different entities such as individuals, enterprises, and governments, as well as multiple types of data. The security risks faced in each stage (such as authorization legality in the collection stage, anti leakage measures in the storage stage, algorithm vulnerability prevention in the usage stage, unauthorized risk control in the sharing stage, and thorough data clearance in the destruction stage) provide a clear framework for technical implementation, helping organizations to choose encryption, access control, desensitization, secure destruction and other technical measures for different stages, and achieve scientific and effective management of the entire lifecycle of data.

2.2 Scope and Type Boundaries of Compliance Data

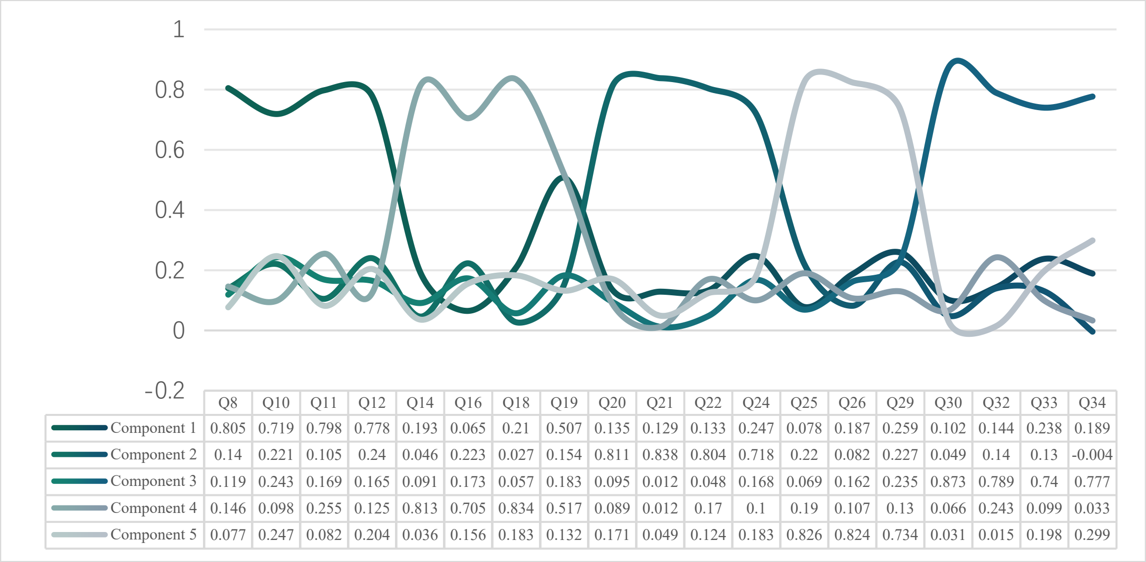
In the digital age, the scope and type boundaries of compliance data are the core issues of privacy and data security research, and their definition needs to be combined with the dual dimensions of personal behavior patterns and corporate governance practices. From a personal perspective, privacy behavior involves users' disclosure of personal information, privacy control applications, and configuration settings. These behaviors not only reflect individual privacy attitudes, but also constitute the basic behavioral patterns of data protection. The privacy paradox reveals the contradiction between users' claims to value privacy but their behavior often violates it. Its formation mechanism can be explained through frameworks such as rational calculation (balancing benefits and risks), social contract (implicit rules between individuals and organizations), and time dimension (cognitive differences between long-term concerns and recent decisions). Individual characteristics (such as gender, age, risk aversion tendencies, privacy self-efficacy), cultural values (collectivism and individualism), institutional environment (laws and regulations, social trust), and technological context (platform design, default settings) collectively shape the complexity of privacy behavior, which in turn affects the specific scope of compliance data - for example, the compliance protection needs of sensitive personal information (such as biometric features, browsing history) are significantly higher than those of non sensitive data. From an enterprise perspective, the type boundaries of compliance data are clearly defined in data lifecycle management. Research on enterprise data security has gradually expanded from technical prevention and control (encryption, blockchain, federated learning) to management frameworks (organizational structure, compliance governance, accountability mechanisms) and value balance (coordination of security and data flow). The entire lifecycle of data is divided into stages such as collection, storage, use, transmission, and destruction, and differentiated security measures need to be implemented in each stage: the collection stage emphasizes data classification and grading (such as grading based on sensitivity), the storage stage relies on encryption technology and trusted backup, the use stage limits the risk of secondary utilization through privacy computing and anonymization technology, and the transmission stage enhances credibility with the help of blockchain. In addition, policies and regulations (such as ISO/IEC 27001) and technical standards (such as the principle of data minimization) further define the type boundaries of compliance data, such as requiring companies to collect only necessary data, restricting the cross-border flow of sensitive data, and clarifying differentiated protection requirements for public data, enterprise data, and personal data. In summary, the scope and type boundaries of compliance data are the result of dynamic balance, which not only needs to respond to individual privacy demands and behavioral conflicts, but also needs to adapt to the evolution of enterprise technology governance and policy norms, ultimately achieved through full lifecycle management, integration of multidimensional influencing factors, and technology institutional collaboration.

3. Research method

3.1 Empirical Study on Cycle Optimization Strategies for Compliance Data Platforms

In the design of the compliance data platform architecture, this study is based on the data lifecycle theory and constructs a privacy concern measurement framework that covers the entire process. Through empirical research, its scientificity and effectiveness have been verified, providing theoretical support and practical basis for platform design. The data lifecycle includes five stages: collection, storage, use, sharing, and destruction. Privacy concerns present differentiated characteristics in each stage: during the collection stage, attention should be paid to users' emotional attitudes towards data collection behavior (such as opposing unauthorized collection) and their right to know needs (such as clarifying the purpose and method of collection), emphasizing users' active control over the data collection process; The storage phase focuses on security risks and quality assurance, requiring an assessment of users' trust in storage security measures (such as anti leakage technology) and the need for continuous monitoring of data accuracy and integrity; During the usage phase, transparency and authorization mechanisms need to be strengthened. Users not only require clear knowledge of the data usage (such as avoiding secondary abuse), but also retain the right to refuse specific uses (such as marketing); During the sharing phase, it is necessary to balance value creation and risk management. Users' concerns about third-party sharing focus on authorization legality (such as not sharing without consent), trust assessment (such as third-party qualifications), and risk awareness (such as the possibility of privacy infringement); The destruction phase, as an easily overlooked aspect, requires additional measurement dimensions such as policy transparency (such as clear retention periods), thorough execution (such as technical verifiability), and user autonomy (such as allowing users to actively trigger destruction). Based on the above framework, an initial quantity containing 27 topic items was designed and measured using a 5-level Likert scale.

To ensure the reliability of the measurement tools, two rounds of questionnaire surveys were conducted in the study. The first survey collected 247 initial questionnaires, and after screening the response time and validation items, 205 valid questionnaires were obtained. Conduct project analysis using the critical ratio method and the overall correlation method, and remove Q13, Q17, Q28, Q31 (critical ratio method, p>0.01) and Q15, Q23, Q27 (overall correlation method), r<0.30), Ultimately, 20 items will be retained; The reliability test showed that the overall Cronbach's alpha was 0.907, and the alpha values of each dimension were greater than 0.7. Only Q9 was removed due to its low CITC value and increased alpha coefficient to 0.869 after deletion; In exploratory factor analysis, KMO=0.895, Bartlett's sphericity test was significant (p<0.001), 5 factors were extracted, and the cumulative variance explained 71.371% (as shown in Figure 1)



	Q8	Q10	Q11	Q12	Q14	Q16	Q18	Q19	Q20	Q21	Q22	Q24	Q25	Q26	Q29	Q30	Q32	Q33	Q34
Component 1	0.805	0.719	0.798	0.778	0.193	0.065	0.21	0.507	0.135	0.129	0.133	0.247	0.078	0.187	0.259	0.102	0.144	0.238	0.189
Component 2	0.14	0.221	0.105	0.24	0.046	0.223	0.027	0.154	0.811	0.838	0.804	0.718	0.22	0.082	0.227	0.049	0.14	0.13	-0.004
Component 3	0.119	0.243	0.169	0.165	0.091	0.173	0.057	0.183	0.095	0.012	0.048	0.168	0.069	0.162	0.235	0.873	0.789	0.74	0.777
Component 4	0.146	0.098	0.255	0.125	0.813	0.705	0.834	0.517	0.089	0.012	0.17	0.1	0.19	0.107	0.13	0.066	0.243	0.099	0.033
Component 5	0.077	0.247	0.082	0.204	0.036	0.156	0.183	0.132	0.171	0.049	0.124	0.183	0.826	0.824	0.734	0.031	0.015	0.198	0.299

Figure 1: Rotation component matrix of the first survey data

After factor rotation, the loadings of all items were greater than 0.7, but there was cross loading in Q19, so it was deleted and 18 items were ultimately retained. The second survey collected 205 valid questionnaires again, with a sample size to measurement item ratio of approximately 11.38:1, covering different gender, age, occupation, and income groups. The reliability test shows that the overall alpha is 0.907, and the alpha of each dimension is greater than 0.7 (as shown in Table1)

Table 1: Reliability statistics of privacy issue dimensions

Dimension	Number of Items	Sample Size	Cronbach's α
Collection	4	205	0.871
Storage	3	205	0.773
Utilization	4	205	0.852
Sharing	3	205	0.824
Destruction	4	205	0.842
Privacy Concern	5	205	0.907

In the validity test, KMO=0.895, Bartlett's test was significant (p<0.001), and the normal distribution test showed that the skewness of each item was between ± 3 and the kurtosis was between ± 10. Confirmatory factor analysis showed that the first-order and second-order models fit well (as shown in Table 1): CMIN/DF=1.383-1.386<3, RMSEA=0.043<0.08, GFI, NFI, IFI, CFI>0.9, PNFI=0.742-0.769>0.5, The second-order model was selected as the final model due to its superior simplicity (with a PNFI of 0.027) and better alignment with privacy concerns as a potential conceptual framework. The reliability and validity test results showed that the combined reliability (CR) of the five dimensions in the first-order model was>0.7, and the average variance extraction (AVE) of the storage, use, and destruction dimensions was slightly lower than 0.6 but>0.5; In the second-order model, the CR of the high-order factor "privacy concerns" is 0.850, and the AVE is 0.534, both of which reach the ideal level. The standardized factor loadings of all items are>0.6 (mostly>0.7), indicating good convergence validity; The correlation coefficient matrix shows that the square roots of AVE in each dimension (collected 0.794, stored 0.726, used 0.769, shared 0.781, destroyed 0.758) are greater than the inter dimensional correlation coefficients, indicating good discriminant validity.

Based on the above analysis, the architecture design of a compliance data platform needs to integrate multiple factors including individuals, enterprises, and governments. Through a dynamic adjustment mechanism of perceived risks and perceived benefits, a privacy protection system covering the entire lifecycle should be constructed. The platform needs to clarify data retention and destruction policies, strengthen technical execution standards (such as thorough data destruction verification) to ensure users' right to know and control; Combining government regulatory requirements (such as the right to be forgotten) with enterprise security management practices (such as protection effectiveness), establish a mechanism to enhance user trust; By measuring users' privacy concerns at different stages of their lifecycle, dynamically adjusting data usage strategies, balancing data value release and privacy protection needs, and ultimately achieving the synergy between user trust enhancement and data value release.

3.2 Data processing flow and task scheduling mechanism

This study focuses on the measurement of privacy concerns and task scheduling mechanisms in the data processing flow, and constructs a research framework for data disclosure willingness based on privacy computing. Through empirical analysis, the impact mechanism of multi-agent factors on user data disclosure behavior is verified. The research hypothesis system takes perceived risk as the core intermediary variable, integrating multiple influencing factors such as personal experience, Internet knowledge, platform reputation, protection effectiveness, legal familiarity and policy effectiveness: privacy concerns (H1) and personal experience (H2) have a significant positive impact on perceived risk; Internet knowledge (H3), platform reputation (H4), protection effectiveness (H5), legal familiarity (H6) and policy effectiveness (H7) have a negative moderating effect on perceived risk. Furthermore, perceived risk (H8) and perceived return (H9a) jointly affect data disclosure willingness, and perceived return is assumed to moderate the negative impact of perceived risk on disclosure willingness (H9b).

The model construction (Figure 2) shows the dynamic relationship among variables

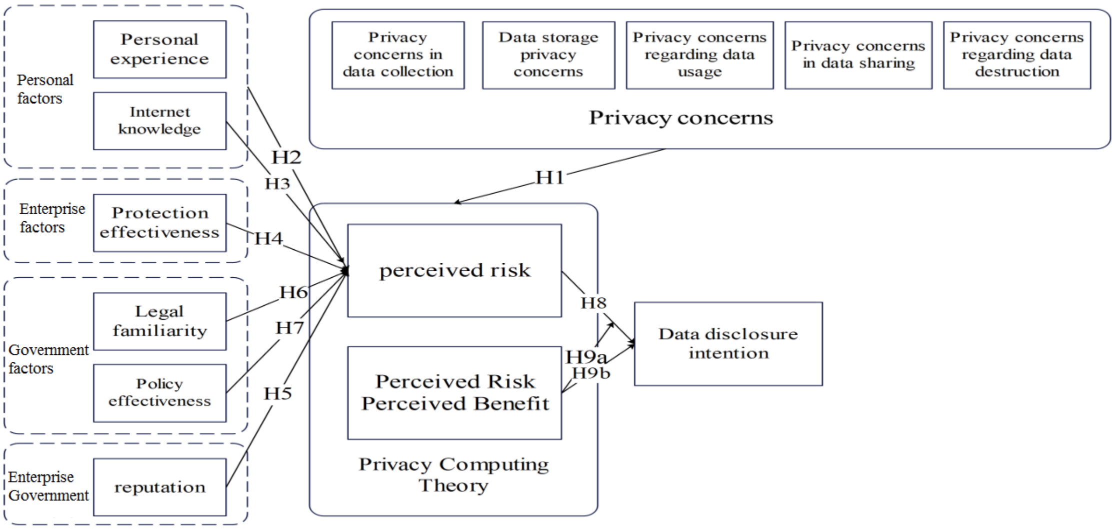


Figure 2: Research Model on Data Disclosure Intention Based on Privacy Calculation

privacy concerns at each stage of the data life cycle (collection, storage, use, sharing, destruction) as a pre variable, through personal experience, Internet knowledge and other subjective factors, combined with external constraints such as platform technology protection (PPP), reputation (REP), legal policy (FGL/REG), jointly affect the user's trade-off between perceived risk (PR) and perceived benefit (PB), and ultimately determine the willingness to disclose data (INT).The research design adopts the questionnaire survey method. The questionnaire contains 52 measurement items, covering privacy concerns (COL/ST/USE/SHA/DEL), personal experience (PE), Internet knowledge (KNOW), protection effectiveness (PPT), reputation (REP), legal familiarity (FGL), policy effectiveness (REG), perceived risk (PR), perceived benefit (PB), and data disclosure willingness (INT) and other potential variables. All items are adjusted based on existing literature and evaluated using the 5-level Likert scale. The third survey collected data through the "Questionnaire Star" platform. There were 512 valid samples. The sample characteristics showed good representativeness: gender balance (51.37% men and 48.63% women), age concentration of 18-45 years old (major Internet users), education level of undergraduate students (47.27%), graduate students and above (21.29%), occupation distribution covering enterprise employees (48.24%), government/public institution employees (19.53%), monthly income concentration in the range of 3000-10000 yuan (67.17%), rich Internet use experience (37.89% users have used the Internet for 3-5 years, 25% users have used it for more than 10 years). The selection of platform types is balanced (33.40% for government platforms, 33.59% for well-known enterprise platforms, and 33.01% for unknown enterprise platforms) to ensure the extrapolation of research conclusions.

3.3 Empirical Study on Data Disclosure under Privacy Calculation

This study focuses on security control and privacy protection mechanisms, constructs a theoretical model of data disclosure willingness, and systematically explores the impact mechanism of multiple subject factors on user data disclosure behavior. The model integrates privacy concerns at all stages of the data life cycle (collection, storage, use, sharing, destruction) as the antecedent variable, combines subjective factors such as personal experience, Internet knowledge, as well as external constraints such as platform technology protection, reputation, legal policies, and influences data disclosure willingness by balancing perceived risks and perceived benefits.

Research proposes a series of hypotheses and verifies that privacy concerns and personal experience have a significant positive impact on perceived risk; Internet knowledge, platform reputation, protection effectiveness, legal familiarity and policy effectiveness have a significant negative moderating effect on perceived risk; The negative impact of perceived risk on data disclosure willingness (coefficient -0.392) and the positive impact of perceived return (coefficient 0.316) alleviate the negative effect of perceived risk, indicating the existence of a risk return trade-off mechanism in user decision-making.The study adopted the questionnaire survey method, designed 52 measurement indicators (including privacy concerns, personal experience, Internet knowledge and other variables), and collected 512 valid samples (5-level Likert scale) through the "Questionnaire Star" platform. The data analysis showed good reliability and validity (Cronbach's alpha>0.7, KMO=0.911, significant Bartlett's test), and the hypothesis was validated by structural equation modeling.Main findings: Privacy concerns have differentiated impacts at various stages of the data lifecycle, with the most significant impact observed during the collection stage; There are significant differences in platform types, with public service platforms having the highest reputation and the lowest perceived risk, while unknown institution platforms have the highest perceived risk. Research suggestion: Enterprises need to strengthen transparent authorization mechanisms and technical protection for data collection; Relevant institutions should improve data security policies, strengthen legal popularization and platform certification; Users enhance their digital literacy to strengthen risk control. Subsequent research will combine evolutionary game models to deepen the structural relationship analysis between enterprise security management capabilities and policy effectiveness.

4. Results and discussion

4.1 Design of Credit Risk Assessment Model for Supply Chain Finance

This study focuses on the frequent occurrence of data breaches and elevates data security from a technical issue to an organizational management level. Based on the theory of data lifecycle (collection, storage, use, sharing, and destruction stages), an enterprise data security management framework is constructed to address the shortcomings of traditional research in analyzing the structural relationships of security management elements. The study adopts a mixed method: firstly, through text encoding of 87 articles from CNKI and Web of Science in the past 5 years (NVivo 14 analysis), 55 initial concepts are extracted and summarized into 13 initial categories, covering technical control points (such as encrypted storage, transmission protocols) and management requirements (such as classification and grading, sharing protocols) at each stage; Secondly, use DEMATEL to quantify the importance of factors and their impact pathways; Finally, the system framework is constructed by combining AISM and MICMAC analysis. The results show that legality and compliance (A11) and data classification and grading (A12) are the core driving factors (high driving low dependence), while the destruction stage factors (A51, A52) belong to low driving high dependence, indicating strong overall system stability.

4.2 Model experiment

This study focuses on compliance behavior recognition and rule construction engine. Based on the data lifecycle theory, the DEMATEL-AISM-MICMAC comprehensive model is used to deeply analyze the 13 key influencing factors of enterprise data security management, and systematically reveal their transmission mechanism and evolution rules. We construct the direct impact matrix, normative impact matrix, and comprehensive impact matrix using the DEMATEL method to quantify the strength of the interaction between various factors. The direct impact matrix is based on pairwise rating data from 15 experts (rating criteria 0-4 points), and the matrix G is normalized by row and maximum values (formula1: $\mathrm{G} = \frac{1}{\mathrm{s}}\mathrm{Z}$, s is the row and maximum values). Then, the comprehensive impact matrix T is calculated through infinite series convergence (formula 2: $\mathrm{T} = \mathrm{G}(1 - \mathrm{g})^{-1}$, I is the identity matrix). The analysis results show that data classification and grading (A12) and legality and compliance (A11) are the core driving factors, with their influence degrees reaching 1.8769 and 1.7233, respectively, and centrality reaching 2.6973 and 2.6916, significantly higher than other factors; The storage security measures (A21) serve as the connection point between technical and management requirements with a centrality of 2.3381, while the physical destruction of media in the data destruction stage (A52) exhibits typical result oriented characteristics (causal degree -0.3857). The causal diagram visually illustrates the transmission relationship between factors: data classification and compliance management form a security control system covering the entire lifecycle by positively influencing storage, processing, and sharing measures at each stage. Furthermore, based on the AISM model for hierarchical analysis, an adjacency matrix (Formula 3) $\begin{cases} A_{IJ} = 1, t_{ij} \geq \lambda \\ A_{IJ} = 0, t_{ij} < \lambda \\ \lambda = \mu + \sigma \end{cases}$ was constructed by setting a threshold (the sum of the average value of the comprehensive influence matrix 0.07156 and the standard deviation 0.03808)And calculate the reachable matrix (determining the factor hierarchy. The hierarchical extraction results show that factors are divided into three levels: the root layer (L3), the intermediate layer (L2, L1), and the execution layer (L0): the root layer includes legality and compliance (A11) and data classification and grading (A12), forming a loop relationship that dominates the entire system; The middle layer covers access control (A31), data processing and anonymization (A32), storage security measures (A21), and shared protocol management (A42), serving as a bridge between the two; The execution layer includes data backup (A22), storage media management (A23), security auditing (A33), transmission security (A41), third-party management (A43), destruction strategy (A51), and media destruction (A52), reflecting specific operational characteristics. The directed topology hierarchy diagram shows that execution layer factors such as transmission security and destruction strategy are directly affected by the root layer, confirming the constraints of the preceding steps on the end measures.

This study analyzes the driver dependency relationship of data security governance through the MICMAC model and constructs a driver dependency distribution map: legitimacy and compliance (A11) and data classification and grading (A12) exhibit high driver low dependency characteristics (driver value 13, dependency value 2), which belong to the core driving factors; The storage security measure (A21) is medium driver high dependency (driver value 4, dependency value 5); The middle layer factors (such as access control A31, shared protocol management A42) are mostly low driver medium dependencies; The destruction stage factors (A51, A52) belong to low drive high dependency. The distribution diagram shows that the system has no fully autonomous or highly interconnected factors, and has strong stability. It is necessary to provide targeted support for dependency groups (A11, A12) to avoid formalization. Combining the DEMATEL-AISM-MICMAC model for system integration (direct impact matrix, reachability matrix, and drive dependency analysis), the high centrality and low dependency of the core driving factors (A11, A12) were verified, confirming their priority optimization status. Through technological protection (such as storage security measures) and collaboration with third-party governance, dynamic security control can be achieved throughout the entire lifecycle, ultimately utilizing rule engines to accurately identify and optimize compliance behavior.

4.3 Effect analysis

This study focuses on the tripartite game mechanism among government, enterprises, and individuals in data security governance. By replicating dynamic equations, it reveals the core impact of regulatory costs (C_g) and policy effectiveness (E.p) on government strategies: when regulatory costs are below a critical value, the government tends to strengthen regulation (1); If the cost exceeds the limit, it will turn to weak regulation (0), leading to governance failure; Policy effectiveness can enhance strong regulatory stability, but high costs can weaken policy sustainability. Further construct a three party game model and simulate a data security warning mechanism: initially set as an ideal equilibrium (0.99, 0.99, 0.99), simulate data leakage events through sudden changes in risk parameters, set warning triggering conditions (personal strategy value<0.6 and lasting for 0.5 time units, or change rate dz/dt<-0.7), covering the entire cycle of risk impact, triggering, intervention, and recovery. Simulation shows that under appropriate parameters, the system can converge to (1,1,1), but there are sensitive parameters (individual perceived risk, compliance cost of enterprises, and government regulatory cost exceeding the threshold will be imbalanced) and critical characteristics. The warning mechanism can effectively identify risks, and the higher the risk intensity, the earlier the warning. However, the intervention window is short and requires quick response. Comparison of intervention plans shows: system crashes without intervention; Mild intervention (only improving corporate reputation) has limited effectiveness; Medium to high-intensity collaborative intervention (synchronously enhancing reputation, compliance investment, and returns, with enterprises intervening first and the government following up) can effectively prevent deterioration and promote recovery. The time for the system to recover to the 0.95 threshold is 8.21 time units earlier than single subject intervention. The study emphasizes that data security requires collaboration among three parties, focusing on critical system characteristics, optimizing policy design (balancing regulatory intensity and cost), establishing a dynamic cost sharing mechanism, and combining real-time warning and multi-party collaborative intervention to provide a theoretical and practical path for compliance dynamic supervision and risk response under big data engineering and intelligent analysis frameworks.

5. Conclusion

This study constructs a multi-agent collaborative data security mechanism analysis framework based on the data lifecycle theory, and conducts systematic research from the micro individual to the macro system level. At the individual level, privacy variables throughout the data lifecycle (collection storage use sharing destruction) are identified through a privacy concern scale, and a seven factor model is constructed that includes privacy concerns, personal experience, network knowledge, protection effectiveness, legal awareness, policy effectiveness, and corporate reputation. It reveals that risk perception has a stronger inhibitory effect on data disclosure than profit incentives, and personal experience has the greatest impact. Protection effectiveness and corporate reputation can reduce risk perception; At the enterprise level, text encoding and DEMATEL methods are used to extract key elements such as data classification and grading, compliance, etc. The AISM model is used to construct a three-layer architecture consisting of the root layer (compliance, data grading), middle layer (access control, desensitization processing), and execution layer (backup management, media control). After MICMAC verifies the rationality of the hierarchy, governance priorities are clarified; At the government level, based on LDA thematic analysis, a three-dimensional framework of "policy tools subject life cycle" was constructed. It was found that there are three major imbalances in current policies: the tool dimension focuses more on the environment/supply type than the demand type, the subject dimension focuses more on government regulation than data rights and professional support, and the cycle dimension focuses more on collection and use than the destruction stage. It was pointed out that the systemic defect of "heavy collection and light destruction" needs to strengthen technical norms and policy support to achieve full cycle closed-loop management; In terms of collaborative mechanisms, a three party game model is used to verify that the government enterprise individual can converge to an ideal equilibrium under appropriate parameters. The simulation of the warning mechanism shows that high-intensity collaborative intervention can improve system stability. Based on this, four-dimensional governance suggestions are proposed, including building a differentiated governance system for the data lifecycle, enhancing individual data security awareness, improving enterprise protection frameworks, reconstructing policy systems, and establishing hierarchical warning and dynamic supervision mechanisms. Although the research has limitations such as insufficient sample representativeness (limited coverage of digital vulnerable groups), theoretical framework scalability (requiring connection between DSMM system and institutional theory), policy analysis scope to be expanded (such as GDPR comparison), and game model optimization space (including decision randomness), it still provides a systematic analysis framework and practical path for multi-agent collaborative data security compliance governance.

References

[1] Lai L. (2025). Data-Driven Credit Risk Assessment and Optimization Strategy Exploration. European Journal of Business, Economics & Management, 1(3), 24-30.

[2] Jadiga, S. (2024). Big Data Engineering Using Hadoop and Cloud (GCP/AZURE) Technologies. International Journal of Computer Trends and Technology, 72(8), 60-69.

[3] Zhu, Z. (2025). Cutting-Edge Challenges and Solutions for the Integration of Vector Database and AI Technology. European Journal of AI, Computing & Informatics, 1(2), 51-57.
[4] Feng, Y., Li, Y., Wang, K., & Liu, L. (2025). A review of the applications of big data and artificial intelligence in oilfield reservoir and fluid dynamics simulation: Feature analysis and development optimization. Advances in Resources Research, 5(1), 46-61.
[5] Jing, X. (2025). Research on the Application of Machine Learning in the Pricing of Cash Deposit Products. European Journal of Business, Economics & Management, 1(2), 150-157.
[6] Davoudian, A., & Liu, M. (2020). Big data systems: A software engineering perspective. ACM Computing Surveys (CSUR), 53(5), 1-39.
[7] Yang D, Liu X. (2025). Collaborative Algorithm for User Trust and Data Security Based on Blockchain and Machine Learning. Procedia Computer Science, 262, 757-765.
[8] Yang, W., Zhang, B., & Wang, J. (2025). Research on AI economic cycle prediction method based on big data[C]//Proceedings of the 2025 International Conference on Digital Economy and Intelligent Computing. 2025: 13-17.
[9] Chen A. (2025). Research on Intelligent Code Search Technology Based on Deep Learning. Pinnacle Academic Press Proceedings Series, 2, 137-143.
[10] Wu X, Bao W. (2025). Research on the Design of a Blockchain Logistics Information Platform Based on Reputation Proof Consensus Algorithm. Procedia Computer Science, 262, 973-981.
[11] Gami, S. J. (2025). Big Data in Smart Learning: Leveraging Data Engineering for Advanced Educational Solutions. Smart Education and Sustainable Learning Environments in Smart Cities. IGI Global Scientific Publishing, 2025: 139-154.