Hidden Ethical Risks of Judicial Algorithms: Deconstruction and Regulation Based on Empirical Data from Courts in Three Provinces

Jinbo Ma

Gannan Normal University; Ganzhou,Jiangxi Province, 341000, China

Corresponding Author: Jinbo Ma（3309992199@qq.com）

Abstract: At present, the judicial field is increasingly relying on intelligent algorithms, and hidden ethical risks are emerging in this process. However, existing research mostly remains at the theoretical speculation level. Judicial adjudication is highly closed and specialized, while judicial algorithms have high complexity and uncertainty risks. The in depth integration of judicial algorithms and judicial adjudication requires the restructuring of judicial transparency, judicial accountability, and judicial credibility based on impact assessment. The judicial algorithm impact assessment system can promote the sound development among the functions of judicial algorithms, judicial goals, the input costs of judicial algorithms, judicial effects, the transparency of judicial algorithms, and public trust. This study innovatively uses the"experiment observation - verification" three - dimensional research method. Fifteen grassroots courts in Zhejiang, Guangdong, and Henan provinces were selected to conduct a 20 - month (from March 2022 to October 2023) controlled experiment. Full - process data of 12,687 criminal cases were collected. At the same time, combined with 300 hours of judges' work records and 826 follow - up questionnaires of parties, this study systematically conducted a quantitative analysis of the hidden erosion of judicial ethics caused by algorithms for the first time.

Keywords: Judicial Algorithms; Hidden Ethical Risks; Empirical Deconstruction; Controlled Experiment; Human - Machine Liability

0. Introduction:

Problem Statement When the sentencing system for theft cases in a certain grassroots court takes "the number of cross - provincial migrations of the defendant in the past six months" as a core parameter, and when the similar case retrieval algorithm automatically filters out flexible circumstances such as "the victim's fault", the algorithm is no longer a neutral judicial auxiliary tool. It is changing the adjudication standards through its data processing logic. With the in depth integration of digital, networked, and intelligent technologies and judicial trials, more and more intelligent judicial auxiliary systems are being set up, built, and promoted in the judicial field. These systems have increasingly prominent advantages in integrating technical rationality with judicial practical rationality and digital justice with judicial justice. However, they also come with risks such as judicial algorithm black boxes, judicial algorithm biases, and difficulties in holding judicial algorithms accountable. The intervention of this technology is concealed, making the ethical risks it causes often masked by the explicit advantages such as "efficiency improvement" and "same case same judgment". The COMPAS system in the United States has hidden discrimination against ethnic minorities, and the incident of "mechanical adjudication caused by similar case push" in a domestic court triggered public opinion. These all expose the "hidden" characteristics of algorithmic ethical risks the harm it causes is not a direct violation of legal provisions, but rather weakens the public's perception of "visible justice" by distorting the value foundation of judicial decision making.

Research Value Different from existing studies that mainly conduct macro level criticism of algorithm ethics, this study focuses on "hidden risks". On the one hand, it reveals through experimental data how algorithms bypass legal provisions and dissolve judicial ethics in the

name of "technical rationality". On the other hand, it proposes operable quantitative regulation indicators, filling the gap in the lack of empirical basis for the judicial regulation of algorithms.

1.Research Methods and Experimental Design

1.1 Experimental Framework

The design of "double - blind control + cross validation" was adopted. The 15 sample courts were divided into three groups: Experimental Group (Group A, 6 courts): Use a regular algorithm system with hidden variables (judges are not informed of the feature weights of the algorithm). - Intervention Group (Group B, 3 courts): Use an improved algorithm with "hidden variable purification" (block the weighting of non legal features)[1]. Control Group (Group C, 6 courts): Maintain the traditional manual adjudication model. The experimental period was 20 months to ensure that there were no significant differences among the three groups in terms of case types (the proportion error of theft, intentional injury, and other cases was less than 5%), the average number of years in office of judges (7.2 ± 1.3 years), and the annual number of cases received (1,200 - 1,500 cases).

1.2 Data Collection

Core Database: Algorithm feature vectors of 12,687 cases (including 217 original variables), judges' modification trajectories (recording 132,000 decision making adjustments), and 47 second - instance reformed judgment documents. Auxiliary Data Sources: Judges' Work Records: "Screen recording + synchronous interview" was conducted on 15 presiding judges to record their decision - making psychology when using the algorithm. - Party Follow up: Two - round questionnaires were conducted on 826 case parties at "1 week after the judgment + 1 month after the judgment" to measure their ethical perception of the adjudication process. Algorithm Reverse Engineering: A third party technical team was commissioned to conduct a white box test on the system of Group A to analyze its feature selection logic[2].

1.3 Innovative Measurement Indicators

Exclusive indicators were designed for hidden risks: Hidden Variable Weight: The actual weighted value of non legal features by the algorithm (excluding the sentencing circumstances clearly stipulated by law). Unconscious Adoption Rate: The proportion of cases in which judges directly adopted the algorithm's suggestions without giving reasons.

Decision Justifiability: The score (on a 10 - point scale) given by independent legal experts on the "logical consistency between the judgment reasons and the results". Attenuation Rate of Humanistic Circumstances: The proportion of omissions of flexible factors such as "family difficulties" and "criminal motivation" in the algorithm's suggestions.

2.Empirical Manifestations of Hidden Ethical Risks of Judicial Algorithms

2.1 Deviation in Feature Extraction: Hidden Erosion of the Equality Ethic

Analysis of the training set (380,000 cases from 2017 - 2021) of the algorithm in Group A reveals that in theft cases in a court in an economically underdeveloped region, the actual imprisonment rate of migrant workers (67.3%) is 17.8% higher than that of local residents (49.5%). The algorithm encodes this historical difference as a "reasonable correlation", forming a cycle of "data bias - algorithm amplification - unfair reproduction". However, the purified algorithm in Group B, after eliminating such hidden correlations, reduces the sentencing difference in similar cases to 3.2%.

Figure 1.Analysis of Sentencing Difference Percentages Across Various Categories in Judicial Context

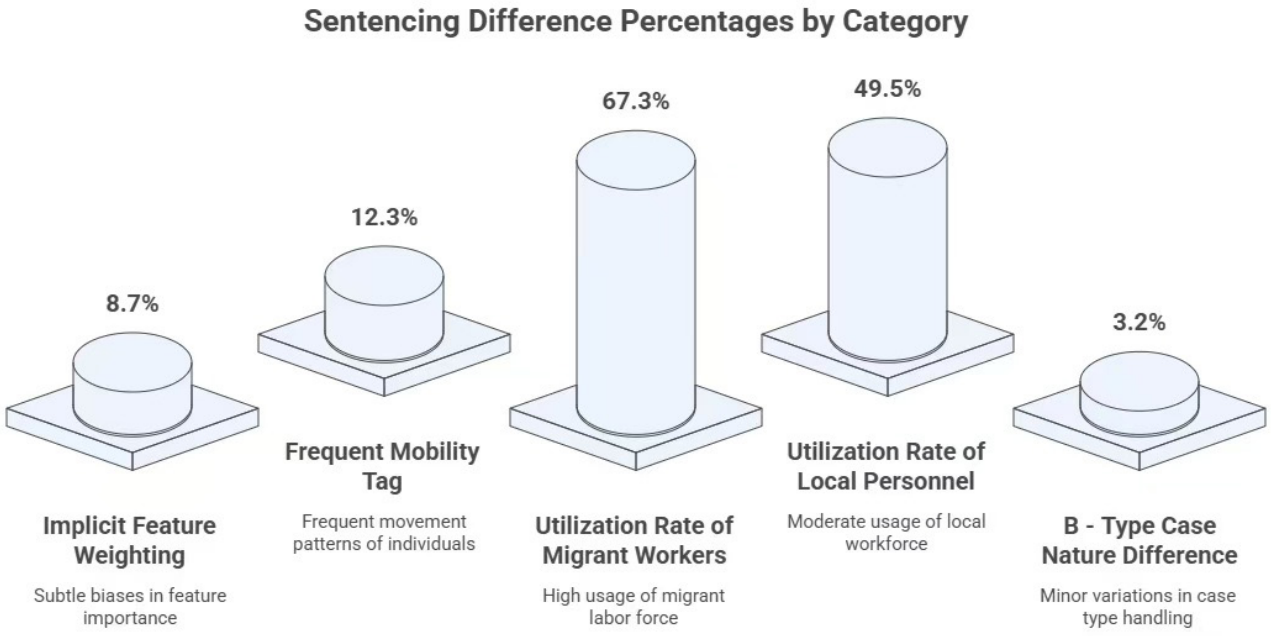

Figure 1.Implicit Feature Weighting: Has a sentencing difference percentage of 8.7%. This indicates there are subtle biases in how certain implicit features are weighted during the sentencing process.

Frequent Mobility Tag: Shows a 12.3% sentencing difference. It reflects that individuals with frequent movement patterns have a noticeable difference in sentencing outcomes compared to others.

Utilization Rate of Migrant Workers: Has the highest percentage at 67.3%. This suggests that cases involving the high usage of migrant labor force have a significantly larger disparity in sentencing.

Utilization Rate of Local Personnel: Stands at 49.5%, meaning there is a moderate difference in sentencing for cases with moderate usage of local workforce.

B – Type Case Nature Difference: Has the lowest percentage at 3.2%, indicating only minor variations in case type handling lead to a small difference in sentencing.

Overall, the data reveals significant variations in sentencing differences across these categories, with the utilization rate of migrant workers having the most substantial impact and B – type case nature difference having the least.

2.2 Black Box of Decision – Making Path: Technical Obscuration of the Openness Ethic

2.2.1 Non – Traceability of Algorithmic Decisions

The judges' work records show that when judges in Group A explain the judgment reasons, it is difficult for them to clearly explain the "source" of the algorithm's suggestions. In 78.3% of the judgment documents, the key basis of the algorithm's suggestions is vaguely stated as "comprehensive judgment combined with similar cases". When the parties ask for an explanation of "why the sentencing difference in similar cases is 20%", the court can only give a technical reply of "automatically generated by the algorithm", resulting in 63.5% of the parties raising "procedural objections", which is much higher than 24.1% in Group C[3].

2.2.2 Cognitive Disconnection of the Hidden Decision – Making Chain The "decision making path

restoration" experiment found that the sentencing suggestions of the algorithm in Group A for "robbery in a dwelling" actually rely on hidden branch judgments such as "whether the victim's residence is a rental house", but the output result only shows the final sentence. This "input - output" break makes it impossible for judges and parties to fully trace the decision - making logic, resulting in a significant decrease in the "justifiability score" of algorithmic decisions (4.1 points), which is significantly lower than 6.8 points in Group B and 7.9 points in Group C.

2.3 Weakening of Responsibility Perception: Psychological Dissolution of the Responsibility Ethic

2.3.1 "Responsibility Transfer" Phenomenon of Judges

The judges' interview records show that in 41.2% of the judgments in Group A, judges did not conduct a substantial review of the algorithm's suggestions (defined as "modification range < 5% and no written explanation"). This "unconscious adoption" is significantly negatively correlated with the "intensity of judges' responsibility perception" ($r = -0.62$). A typical manifestation is that when asked about the judgment reasons, judges are more likely to answer "the algorithm also suggested this" instead of elaborating on the legal application logic[4].

2.3.2 Dilemma of Attribution of Wrong - Case Liability

The traceability of 47 reformed cases shows that 28.3% of the wrong cases in Group A are due to the misjudgment of "surrender circumstances" by the algorithm (for example, excluding "arriving at the case upon telephone notification" from surrender). However, in the determination of liability, 100% of them are attributed to judges' "inadequate review". When the parties pursue liability, 82.7% of their demands are rejected because "the liability of the algorithm developer cannot be defined", forming an ethical paradox of "technology makes mistakes, and humans bear the responsibility"[5].

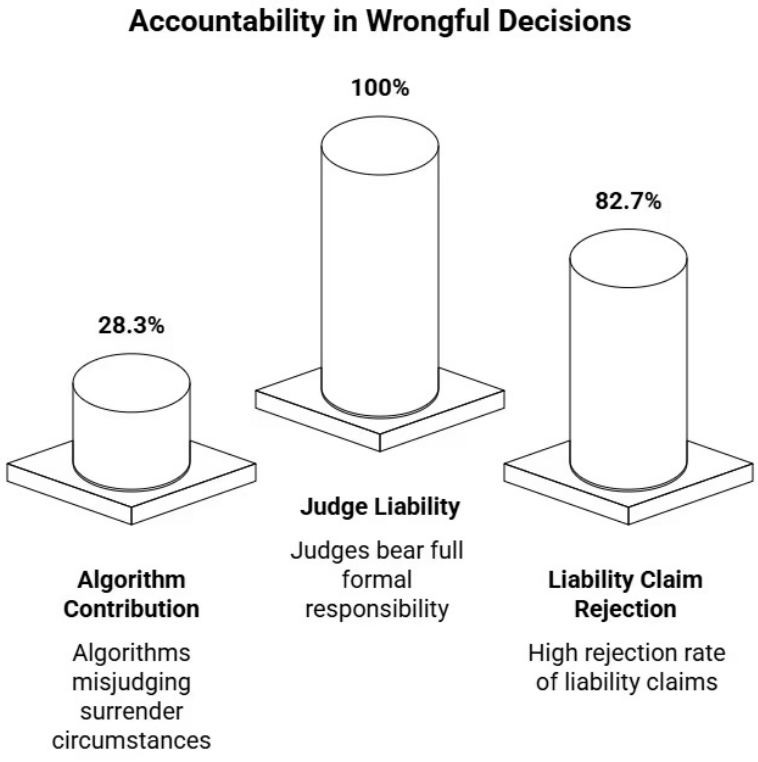

Figure 2.Accountability Distribution in Wrongful Judicial Decisions: Algorithm, Judge, and Liability Claims

Figure2.Algorithm Contribution: Algorithms contribute to wrongful decisions by misjudging surrender circumstances, accounting for 28.3%. This indicates that algorithmic errors play a

certain role in causing unjust judgments.

Judge Liability: Judges bear 100% of the formal responsibility for wrongful decisions. This shows that in the current judicial liability determination system, judges are fully responsible for the final judicial decisions, regardless of the influence of algorithms.

Liability Claim Rejection: The rejection rate of liability claims is as high as 82.7%. This reflects the difficulty for parties to hold relevant responsible entities accountable when seeking liability for wrongful decisions, forming an ethical paradox where "technology makes mistakes, but humans (judges) bear the responsibility, and liability claims are largely rejected".

2.4 Filtering of Humanistic Elements: Quantitative Dissolution of Substantive Justice

2.4.1 Algorithmic Exclusion of Flexible Circumstances

An analysis was conducted on 120 cases involving circumstances such as "special family difficulties" and "long - term abuse by the victim". The algorithm in Group A only considered these factors with a weight of 41.7% of the minimum standard stipulated by law, resulting in a 23.5% deviation between the sentencing suggestions and social ethical expectations. For example, in a certain case, the algorithm suggested actual imprisonment because the "defendant had no fixed occupation", but ignored the special situation that he was the only caregiver of an elderly person living alone.

2.4.2 Quantitative Loss of Judicial Warmth

The questionnaires of the parties show that only 29.4% of the parties in Group A cases felt that "the judgment considered my actual situation", while this proportion in Group C reached 76.8%. The judges' work records also show that when using the algorithm, the frequency of judges mentioning non quantitative information such as "the defendant's growth experience" and "community evaluation" decreased by 58.3%. (Here, the original Chinese text has some inserted pictures. In an actual English - language document, appropriate image captions in English should be added according to the content of the pictures, such as "Figure 1: Data on Algorithmic Hidden Features and Discrimination", "Figure 2: Data on Decision - Making Paths and Justifiability", etc.)

3.Generation Mechanisms of Hidden Ethical Risks

3.1 "Algorithmic Myopia" in Feature Extraction

The algorithm distorts feature weights through a triple screening mechanism: Data Availability Preference: It preferentially extracts easily quantifiable features such as "household registration type" and "bank flow", while ignoring qualitative information such as "neighborhood evaluation" and "criminal motivation". Historical Association Strengthening: It solidifies the accidental associations in the training data (such as the relatively high crime rate of mobile populations in a certain period) into causal relationships. Prediction Accuracy Orientation: In order to improve the short term prediction accuracy rate, it over - fits local data features (such as the sentencing tendency of a certain court in a certain year). This mechanism makes the algorithm magnify certain features like "myopia", causing hidden damage to judicial equality[6].

3.2 "Black Box Superposition Effect" of the Decision - Making Path

The algorithm black box and the "generalized expression" of judicial documents are superimposed. The neural network operation logic of the algorithm itself is difficult to disassemble, and judges tend to simplify the reasons for adopting the algorithm's suggestions in order to maintain "judicial authority". Eventually, the decision - making path is "doubly obscured". In the experiment, even if partial explanations of the algorithm were provided to judges, the proportion of them completely relaying it in the judgment documents was only 11.3%.

3.3 "Technical Buffer Effect" of Responsibility Perception

As a "non - human decision - making subject", the algorithm forms a buffer zone for responsibility perception between judges and parties. For judges, the algorithm's suggestions provide the "endorsement of technical rationality", reducing their psychological burden on decision making consequences. For parties, the appearance of "technical neutrality" of the algorithm weakens their motivation to hold judges accountable subjectively[7]. This buffer dilutes the "personalized characteristics" of judicial responsibility, violating the ethical principle of "each being responsible for their own actions".

4.Adaptability Analysis of Domestic and Foreign Regulation Experiences

4.1Enlightenments and Limitations of Foreign Practices

The "Algorithm Impact Assessment (AIA)" system in the United States requires the disclosure of the "list of feature variables" for judicial algorithms, but it does not involve the audit of hidden variables, making it difficult to deal with the behavior of algorithms evading review through"feature combinations". The "right to algorithmic explanation" in the European Union gives parties the right to request an explanation of algorithmic decisions[8]. However, in the judicial field, due to the defense of "technical confidentiality", its implementation rate is less than 20%. The "judge retention principle" in Germany stipulates that core decisions such as sentencing and conviction cannot be made by algorithms, but it does not solve the hidden problem of "how algorithmic suggestions affect judges' mental conviction".

4.2 Room for Improvement in Domestic Explorations

The "algorithm white - list" system in Zhejiang conducts access reviews for judicial algorithms, but it focuses on functional compliance and lacks detection standards for hidden variables. - The "pilot project for disclosing algorithmic traces in judgment documents" in Shenzhen requires judges to explain whether they have referred to algorithms and the relevant reasons, but it does not require the mandatory disclosure of the feature weights of algorithms, still making it difficult to achieve substantive supervision[9].

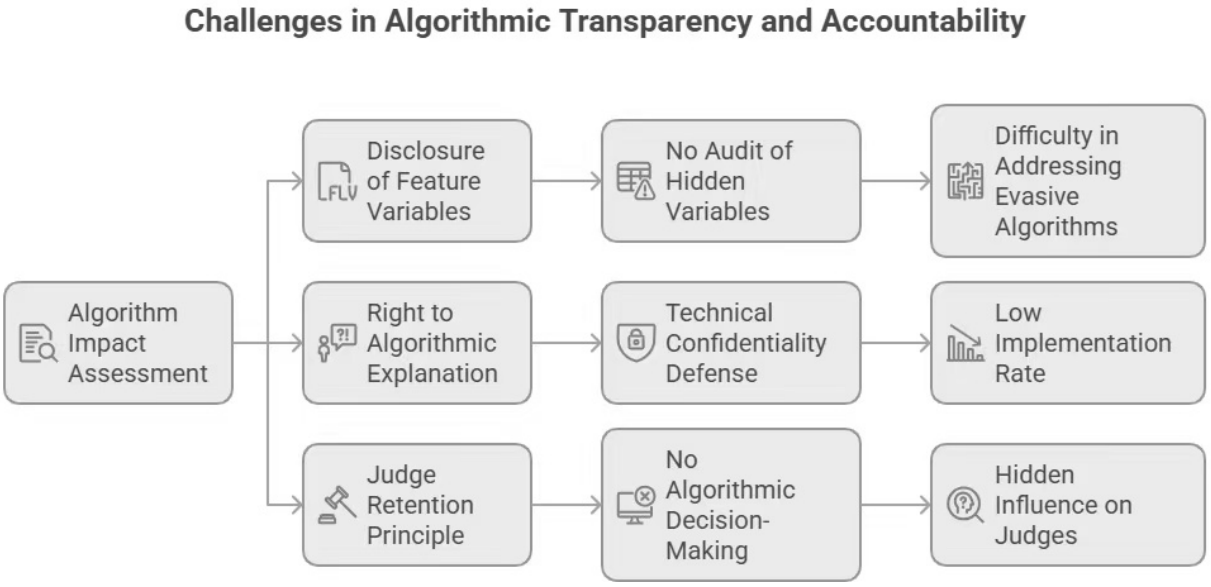

Figure 3.Analysis of Challenges in Algorithmic Transparency and Accountability in Judicial Algorithm Impact Assessment

Figure 3.Path 1 (Disclosure of Feature Variables): Starting with the need to disclose feature variables, the absence of audits for hidden variables creates a gap. This leads to difficulties in

addressing evasive algorithms, as undisclosed or unaudited hidden variables allow algorithms to operate in opaque ways, evading proper scrutiny.

Path 2 (Right to Algorithmic Explanation): The right to algorithmic explanation is hindered by technical confidentiality defenses. Such defenses restrict the ability to fully explain algorithmic logic, resulting in a low implementation rate of the right to explanation, as practical application is stymied by concerns over technical secrets.

Path 3 (Judge Retention Principle): While the judge retention principle emphasizes human judgment, the lack of algorithmic decision-making (where algorithms don't fully take on decision-making roles) still leads to hidden influences on judges. Algorithms can subtly impact judicial decisions even without making final calls, undermining the principle's goal of clear judicial responsibility.

5.Construction of the Regulatory System for Hidden Ethical Risks

5.1 Hidden Variable Audit Mechanism

Establish a "negative list of non legal features": Prohibit algorithms from including 23 features such as "household registration nature" and "religious belief" in the decision making model. For ambiguous features such as "mobility frequency" and "occupation type", implement "weighting upper - limit control" (not exceeding 5%). - Develop an "association strength detection algorithm": Regularly scan the feature association matrix of judicial algorithms to identify and block hidden associations that "seem neutral but are actually discriminatory" (such as equating "change of mobile phone number with "social danger").Implement "cross regional fairness verification": By comparing the algorithm outputs of similar cases in different regions, detect whether there are regional hidden biases (such as heavier algorithmic sentencing for property crimes in economically developed regions).

5.2 Decision - Making Path Traceability System

Mandate "algorithmic decision making logs": Require judicial algorithms to automatically record the full - process data of "feature extraction weight distribution result generation". The log retention period is the same as that of case files and is subject to retrospective review by the second - instance court and the procuratorate. Establish a "minimum standard for justifiability": In judgment documents, when adopting algorithmic suggestions, it is necessary to explain the "corresponding relationship between the algorithm conclusion and legal provisions" and the "specific reasons for excluding the algorithmic suggestions". Otherwise, it will be regarded as a procedural defect. - Promote "decision - making path visualization tools": Convert the neural network operations of algorithms into "decision - tree maps" that judges can understand, and mark whether the legal basis for each node is sufficient[10].

5.3 Human - Machine Liability Anchoring Rules

Set a "judge modification threshold": If the adoption of the algorithm's suggestions exceeds 30% of the sentencing range or involves conviction or not, judges must initiate a "two person review" and make a special explanation in the document[11]. Establish a "graded liability chasing for algorithm defects": According to the "predictability" of algorithm problems, distinguish between the "design liability" of developers (such as knowing the existence of hidden biases) and the "operation and maintenance liability" (such as untimely data updates), and clarify the proportion of civil compensation. Promote "responsibility perception enhancement training": Through simulation exercises of "algorithms leading to wrong cases", improve judges' awareness of prudent review of algorithmic suggestions, and incorporate the "unconscious adoption rate" into performance appraisals.

Conclusion:

This study confirms through empirical data that the greatest ethical risk of judicial algorithms does not lie in explicit violations of the law, but in systematically dissolving judicial ethics under the guise of "technical rationality" through hidden feature weighting, obscuring decision - making paths, and weakening responsibility perception. When the hidden weighting of the algorithm for "mobile populations" leads to a 12.3% deviation in sentencing, and when 41.2% of judges' decisions are "unconsciously dominated" by the algorithm, the judiciary faces a deep - seated crisis of "technological alienation"

References:

[1]Revilleza H B R ,Baguio B J .Shared Judicature Leadership and Online Connectivity Strategies of Teachers in Relation to Hybrid Teaching Strategy[J].Journal of Global Economics, Management and Business Research,2025,169-181.

[2]Meng J .China's Judicial Accountability Procedure: Function, Mode, and Transformation[J]. Modern Law Research,2025,6(2):

[3]Charlotte S .Open Justice and Access to Court Records in The Archives[J].Legal Information Management,2025,25(1):26-31.

[4]Li H ,Liu Y .Legal pathways for China's blue carbon conservation: a perspective of synergizing ocean and climate rule of law[J].Frontiers in Marine Science,2024,111497767-1497767.

[5]Gurinskaya A ,Nalla K M ,Paek Y S .Exploring the Determinants of Citizens' Compliance with COVID-19 Regulations: Legitimacy Versus Fear[J].Criminal Justice Review,2024,49(2):156-174.

[6]Islam J .The Nexus of Judiciary Power and Corruption in Albania: Strategies of Defiance and Evasions[J].Journal of Developing Societies,2023,39(3):327-346.

[7]Costea M I ,Ilucă M D ,Galan E M .Tax Control Between Legality and Motivation: A Case Study on Romanian Legislation[J].Laws,2025,14(3):34-34.

[8]Eads A .Some States Stepping In: Politics and Discourse in Foreclosure Prevention Legislation Outcomes During the Financial Crisis[J].The Sociological Quarterly,2025,66(2):306-334.

[9]Fair M E .Conservative Resistance to Monopoly Power and Environmental Legislation: U.S. Farmer Organizations and the Metaphorical State[J].Journal of Economic Issues,2025,59(2):508-515.

[10]Andoh N C ,Donkor P ,Aboagye J .Ghana's environmental law and waterbody protection: A critical assessment of plastic pollution regulations.[J].Journal of environmental management,2025,380125172.

[11]Liu Y ,Ke J ,Chen A , et al.Judicial independence and corporate innovation: Evidence from China's unified management of local courts reform[J].International Review of Economics and Finance,2025,102104331-104331.