Research on the Judicial Fairness Challenges and Legal Regulation of AI-Assisted Judgments in the Context of Digital Justice

Qinglin Zeng
[1]Party School of Sichuan Committee of C.P.C, Chengdu 610000, Sichuan, China
Corresponding author: Qinglin Zeng (e-mail: 13114436912@163.com)

ABSTRACT Artificial intelligence has become a critical enabling technology for intelligent decision-support systems, creating new opportunities and challenges for trustworthy information processing in digital governance environments. This study investigates the fairness risks and regulatory mechanisms of AI-assisted judicial decision-making by establishing an integrated framework that combines algorithm interpretability, data governance, bias mitigation, and human–machine collaborative control. The proposed framework systematically analyzes the impacts of black-box reasoning, training-data bias, and excessive technological dependence on decision reliability and procedural fairness. To address these challenges, explainable reasoning standards, full-lifecycle data governance strategies, algorithmic fairness auditing, and human-supervised decision protocols are incorporated into a unified regulatory architecture. Furthermore, a traceable reasoning mechanism and adaptive oversight framework are introduced to improve transparency, accountability, and operational robustness in AI-assisted decision systems. The proposed methodology provides practical guidance for explainable intelligent systems, trustworthy information processing, adaptive decision support, and distributed human–AI collaboration, offering potential references for intelligent sensing, secure information management, and next-generation digital service infrastructures.

INDEX TERMS explainable artificial intelligence, trustworthy information processing, human–AI collaboration, adaptive decision support

I. INTRODUCTION

IN recent years, the rapid development of artificial intelligence technology has brought profound changes to both the judicial field and the manufacturing industry, reshaping traditional workflows through automated reasoning and high-speed fabric defect inspection. This technological convergence drives a systemic shift toward intelligent decision-making in legal proceedings and digitized quality control in spinning and weaving mills. Intelligent trial assistance systems have been gradually and widely applied in areas such as sentencing reference and judgment document generation. However, while technology brings efficiency benefits, it also raises deep concerns about algorithmic transparency, data fairness, and discretionary independence. How to safeguard the bottom line of judicial fairness in the process of promoting digital justice has become an important issue of common concern in legal theory and practice.

II. THE SIGNIFICANCE OF AI-ASSISTED JUDGMENT IN THE CONTEXT OF DIGITAL JUSTICE

A. IMPROVING JUDICIAL EFFICIENCY AND CONSISTENCY OF JUDGMENTS

Artificial intelligence-assisted adjudication systems can quickly retrieve and intelligently compare a large number of cases, greatly saving judges the time spent collecting information and reviewing precedents [1]. When AI systems handle the same type of cases, they can rely on a unified data model to generate reference opinions, which is conducive to overcoming the differences in adjudication results caused by different judges' different experiences. This improves the predictability of adjudication and the consistency of legal application from an institutional perspective, and achieves judicial fairness in a wider range [2].

B. PROMOTING THE BALANCED ALLOCATION OF JUDICIAL RESOURCES AND UNIVERSAL ACCESS TO JUSTICE

For a long time, underdeveloped areas have suffered from a lack of judicial resources and legal aid, which has affected the realization of judicial fairness. When the AI-assisted judgment system plays its role, it can overcome

geographical limitations and use technology to deliver high-quality judicial resources to grassroots courts and remote areas, so that local judges can receive assistance from high-quality data models when facing complex cases [3]. This mechanism effectively alleviates the problem of insufficient grassroots judicial capacity, enables judicial resources to be allocated reasonably, and allows judicial fairness to benefit more people in urban and rural areas and regions, thus realizing universal judicial benefits [4].

C. FACILITATING JUDICIAL TRANSPARENCY AND STANDARDIZATION OF JUDGMENTS

The AI-assisted judgment system relies on a standard legal document generation module to assist judges in organizing reasoning according to a standardized logical structure. Concurrently, in the industrial domain, AI-driven quality reporting engines assist inspectors by providing a uniform classification framework for fiber morphology and yarn distribution. While serving different ontologies, these systems respectively reduce arbitrary expressions in judicial rulings and eliminate inconsistent defect classifications in high-speed weaving. The transparency of judicial documents is improved through the standardization of legal reasoning, whereas the accuracy of technical logs is enhanced through digital feature extraction; however, these remain distinct data repositories serving their respective professional fields. The AI system systematically organizes and saves case data, which is conducive to forming a more complete judicial open archive. Parties and the public can easily access the basis of judgments and reasoning processes, thereby improving the visibility of judicial operations. However, if the archived data lacks explainability, the "black-box" nature of the output may conversely erode public trust. Therefore, the integration of interpretable metadata into public archives is essential for providing strong technical support for enhancing judicial credibility [5]. The positive value of AI-assisted judgment can be understood from three aspects: efficiency improvement (addressing backlog volume), resource balance, and transparency and standardization (ensuring predictability and explainable credibility) [6]. Its internal logic is shown in Figure 1.

III. CHALLENGES TO JUDICIAL FAIRNESS POSED BY AI-ASSISTED JUDGMENTS

A. THE BLACK BOX OF ALGORITHMS LEADS TO A LACK OF TRANSPARENCY IN THE REASONING OF JUDGES

Most of the current mainstream AI-assisted judgment systems adopt deep learning architecture, and the internal calculation process is very complicated. Even the system developers find it difficult to give a complete logical explanation for each output result. This technical limitation necessitates a transition from "post-hoc explanation" to "traceability by design." If developers cannot provide an interpretive framework, the system fails to meet the basic legal threshold for judicial application. When the AI system provides sentencing suggestions or judgment references to judges, the system cannot clearly present the reasoning chain. Judges can only accept the conclusive output and cannot trace its formation basis [7]. This kind of "only seeing the result and not the process" operation mode directly impacts the basic requirement that judicial judgments must be based on sufficient reasoning, and puts the argumentation logic in the judgment documents in danger of becoming empty. The parties' right to know and right to question are also substantially infringed, and the legitimacy of the judicial procedure faces a severe test [8].

B. TRAINING DATA BIAS MAY LEAD TO LEGAL DISCRIMINATION RISKS

The quality of AI-assisted judgment systems largely depends on the quality and representativeness of training data. However, historical judicial data itself contains institutional defects and social biases of a certain period. Once records of unfair judgments against a certain group are included in the training set, they will be learned and reproduced by the system in a covert manner [9]. Especially in cases related to sensitive variables such as the parties' regional background, social identity, and ethnic characteristics, data bias will cause the AI system to give different reference conclusions to parties from different groups in different types of cases, thus generating implicit discrimination at the algorithm level. The risk of discrimination cannot be detected by the naked eye and will continue to increase with the large-scale promotion of the system, posing a deep threat to the principle of judicial equality [10].

C. TECHNOLOGICAL DEPENDENCE ERODES JUDICIAL DISCRETION

The independent exercise of discretion by judges is the institutional foundation of judicial fairness. However, as AI-assisted judgment systems are increasingly embedded in judges' practice, some judges will develop path dependence and regard AI suggestions as the substantive basis of judgments rather than as auxiliary references, thereby weakening their initiative in independent judgment. This is even more serious in grassroots courts because the cases are complex. In order to avoid personal responsibility and reduce work pressure, judges follow the system's direction, which makes the power of judicial judgment essentially shift from judges to the technical system. The weakening of judges' subjectivity not only violates the judicial ethics of "people-oriented", but also makes the accountability mechanism unclear. The humanistic background of judicial fairness may be obscured by technological rationality [11]. The above three problems are not isolated, but are layered according to the path of technical defects, transmission mechanism, rights infringement, and systemic risk [12]. Its transmission logic is shown in Figure 2.

IV. LEGAL REGULATORY STRATEGIES FOR AI-ASSISTED JUDGMENTS

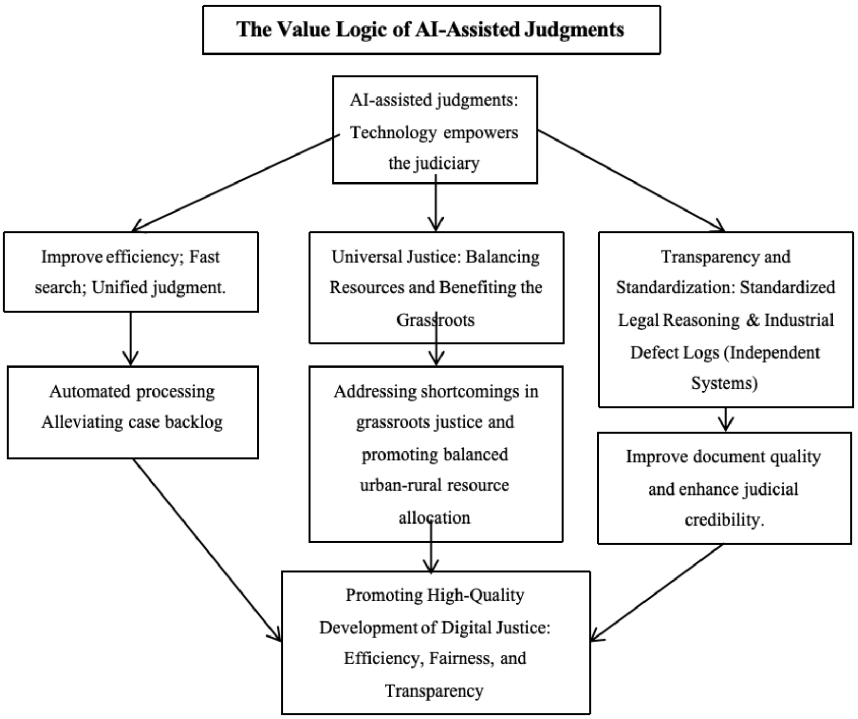

FIGURE 1. The Three-Dimensional Value Logic of AI-Assisted Judgment

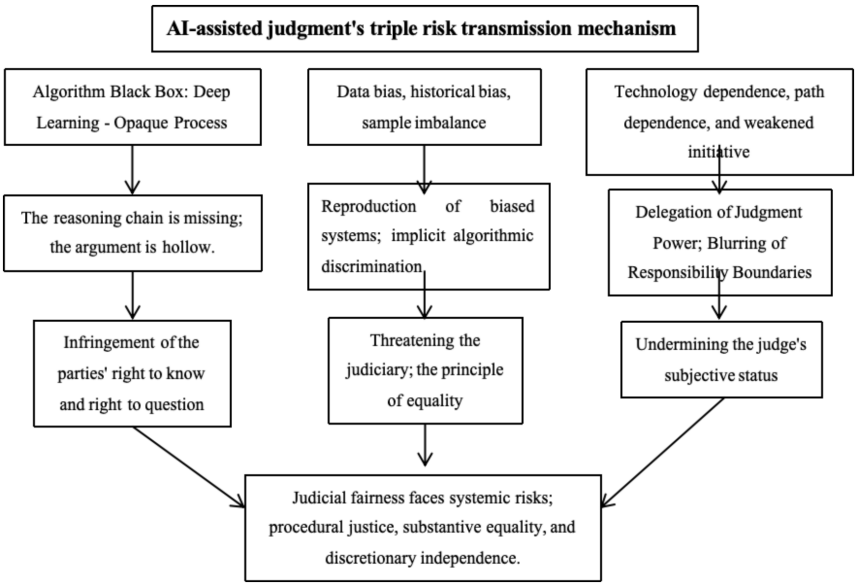

FIGURE 2. Risk Transmission Mechanism of the Three Challenges of AI-Assisted Judgment

A. ESTABLISHING ALGORITHM INTERPRETABILITY STANDARDS: OVERCOMING THE DILEMMA OF REFEREE TRANSPARENCY

For the problem of opaque reasoning in judgments caused by the black box of algorithms, the main regulatory path is to establish the interpretability standard of AI-assisted judgment system by legislation. Specifically, legislation must define the "Minimum Explainability Requirement" (MER) that any commercial or proprietary system must meet. On the one hand, it should be clearly required through judicial administrative regulations that developers—not the state—are responsible for demonstrating the technical traceability of their models. AI systems that remain a "black box" to their own creators shall be legally barred from judicial deployment, thereby placing the technical burden on the industry to innovate transparent architectures. Each time the system gives a reference suggestion, it should provide the judge with the relevant reasoning node explanation so that the judge can understand the reasoning basis given by the system, rather than just receiving the conclusive instructions.

TABLE 1. Three-layer framework for algorithm interpretability regulation

Regulatory level	Core Requirements	Specific measures	Responsible Entity
Legislative normative level	Mandatory requirement for traceability of reasoning paths	Judicial and administrative regulations clearly define interpretability standards; AI systems must output explanations of key reasoning nodes.	Legislative and judicial administrative organs (for setting legal mandates); System Developers (for technical compliance and proof of interpretability)
Technical certification layer	Explainability by Design certification. Pre-launch third-party assessment of the developer's internal logic documentation.	An independent third-party technical body; products that fail to pass certification cannot be used in judicial practice.	Third-party certification bodies, technology developers
Program protection layer	Granting parties the right to know and the right to object	The court document should specify the use of AI; the parties may request an explanation of the basis for the formation of the AI's reference opinions.	Courts, judges

On the other hand, an interpretability certification system should be established before the AI system goes online. Independent third-party technical institutions should assess the transparency level of the system, and those that do not meet the requirements should not be put into use. In addition, the parties should be given the right to know and the right to object. During the trial, the court should be required to explain the basis for the formation of AI-assisted opinions and to indicate the use of AI system in the judgment document. The loophole of empty reasoning in judgments should be blocked from the system and the inherent requirements of judicial procedure justice should be guaranteed [13]. The regulation of algorithm interpretability can be implemented from the following three aspects, as shown in Table 1.

B. IMPROVE DATA GOVERNANCE AND BIAS REVIEW MECHANISMS: PREVENT THE RISK OF JUDICIAL DISCRIMINATION

In order to overcome the bias in the training data, it is necessary to create a systematic governance framework from the perspective of the entire data life cycle. The judicial training data access standards must be stipulated, and the historical case data of specific persons with discriminatory judgment results should be screened for bias, and samples containing discriminatory judgment results should be removed to prevent the training model from solidifying and perpetuating historical biases and to prevent the phenomenon of continuation and solidification during the training process. In the system operation stage, a regular bias audit system should

be established, and the professional evaluation team of the highest judicial administration organ should be granted to conduct periodic fairness tests on the existing AI-assisted judgment system, mainly to examine whether there are system differences in the handling of cases by parties with different social backgrounds. At the same time, it is also necessary to promote diversified data governance, encourage the inclusion of data from different regions, different periods, and different types of cases into the training set, thereby improving the representativeness and inclusiveness of data samples. From the perspective of legal responsibility, the joint responsibility boundary between AI system developers and judicial organs when data bias leads to wrong judgments should be determined, and the data governance level should be continuously improved by relying on the responsibility-driven mechanism, thereby providing an institutional foundation for building a non-discriminatory digital judicial environment [14]. The data governance framework should run through the entire data life cycle, and the regulatory focus and responsible subject of each stage are shown in Table 2.

TABLE 2. Full-cycle regulatory framework for data governance and deviation review mechanisms

Governance	Core Objectives	Regulatory measures	Supporting mechanisms
Data collection	Preventing historical biases from becoming entrenched	Establish standards for training data access; conduct bias screening on historical cases involving specific groups to eliminate discriminatory judging samples.	Data Quality Review Committee
System operation	Timely detection of systemic differences in output	Establish a regular deviation audit system; a professional team conducts periodic fairness tests on the AI system, focusing on detecting differences in output from parties with different backgrounds.	The highest judicial and administrative organ's evaluation team
Data update	Enhancing sample representativeness and inclusivity	Promote diversified data governance; include data from different regions, periods, and types of cases in a balanced manner.	Data management departments, courts in various regions
Legal liability	Forming a data governance incentive mechanism	Clearly define the shared responsibility of developers and judicial authorities when data discrepancies lead to errors; this accountability mechanism drives continuous improvement in data governance.	Legislative bodies, judicial bodies, and developers

C. ESTABLISHING BOUNDARY RULES FOR "HUMAN-MACHINE COLLABORATION": SAFEGUARDING JUDGES' INDEPENDENT JUDGMENT RIGHTS

To ensure that judges' independent discretionary power is not affected by technological dependence, it is fundamentally necessary to clarify the respective roles of humans and machines in judicial judgment through institutional means. Relevant judicial interpretations or technical specifications should clearly stipulate the

basic principle of AI assistance and judge-led decision-making, requiring judges to conduct a "Two-Way Independent Argumentation." Instead of providing a post-hoc justification for adopting AI suggestions, judges must first independently formulate a draft reasoning chain before accessing the AI's recommendation. To counteract path dependence, a "Substantive Deviation Review" mechanism should be implemented. If a judge chooses to adopt an AI suggestion that aligns with a high-probability sentencing model, the judge is required to provide a "Non-Automated Rationale," specifically highlighting the unique humanistic elements or legal nuances of the case that the algorithmic model cannot capture. This shifts the procedural burden from a mere explanation of the result to a demonstration of active cognitive engagement, thereby preventing the rubberstamping of algorithmic outputs. A log-keeping system for AI-assisted judgments should be established to record and preserve information such as the extent to which judges use AI systems. To transform this log from a descriptive tool into a regulatory mechanism, specific

TABLE 3. Institutional Construction Dimensions of "Human-Machine Collaboration" Boundary Rules

Building Dimensions	Core Principles	Institutional Arrangements	Applicable Scenarios
Institutional norms	AI-assisted, judge-led	Mandate "Pre-AI Independent Reasoning" and "Non-Automated Rationale" requirements; judicial interpretations should require evidence of a judge's active modification or critical engagement with the AI suggestion to validate independent discretion.	All types of cases
Oversight mechanism	Behavioral traceability and over-reliance are traceable.	Establish a log system with quantitative alert metrics (e.g., consistency rate thresholds and reasoning divergence requirements); fully record the frequency, content, and correlation of judges' use of AI with the judgment results to flag statistical anomalies of over-reliance.	Routine judicial supervision
Capacity building	Critical evaluation algorithm recommendations	Incorporate AI literacy education into judges' professional training; focus on cultivating the ability to identify algorithmic limitations and critically evaluate AI recommendations.	Judge training system
Mandatory review	Preventing technological overreach	High-risk cases are subject to mandatory manual review; serious criminal cases and cases involving vulnerable groups must be discussed collectively by the collegial panel before AI opinions can be considered.	Serious criminal cases and cases involving vulnerable groups

quantitative thresholds for "over-reliance" must be defined. For instance, an "Alert Trigger" should be activated if the consistency rate between a judge's final ruling and the AI's initial suggestion exceeds a predefined statistical variance (e.g., >90% across nonroutine cases), or if the judge's modification of the AI-generated reasoning nodes is less than a minimum qualitative threshold (e.g., <15% word-count divergence). This provides actionable data support for tracking whether judges rely too heavily on AI in their trials, ensuring the system acts as a functional regulatory firewall rather than a mere record-keeper. Regarding training, AI literacy education should be incorporated into judges' professional training, focusing on cultivating judges' ability to identify algorithmic flaws and critically evaluate AI suggestions. This training should specifically include "Contradictory Case Simulations," where judges practice overriding biased or flawed AI outputs to strengthen their psychological resistance to path dependence, ensuring that technological empowerment does not lead to cognitive capture. The mandatory rule for high-risk cases should be set to manual review. Before making a judgment on cases involving serious crimes or vulnerable groups, the opinions put forward by AI must be collectively discussed by the collegial panel before they can be cited. This metric-driven oversight will build a solid defense against the overreach of technology from the institutional level and safeguard the basic principle of judicial fairness by providing objective benchmarks for judicial accountability. By ensuring that the "black box" is replaced by a transparent reasoning process, the system transitions from an opaque risk to a verifiable tool, thereby resolving the inherent contradiction between algorithmic complexity and judicial accountability. [15] The boundary rules of "human-machine collaboration" can be systematically constructed from three dimensions: institutional norms, supervision mechanisms and capacity building, as shown in Table 3.

V. CONCLUSION

The advent of the digital justice era and the high-speed manufacturing revolution has provided unprecedented technological conditions for improving judicial efficiency and quality monitoring, making the positive value of AI-assisted judgments and automated fabric inspection undeniable, provided that the technical "black box" is effectively regulated through transparency standards. This technological parallelism ensures that both legal processes and spinning operations benefit from enhanced accessibility and data-driven precision within their specific domains. By maintaining the ontological separation between judicial reasoning and industrial inspection data, the system ensures that judicial integrity is not conflated with technical manufacturing metrics. However, systemic risks such as algorithmic black boxes, data bias, and technological dependence cannot be ignored. The realization of judicial fairness fundamentally relies on transparent, equal, and independent institutional guarantees. When AI-assisted judgments are proposed, legal regulation should follow suit, constructing a comprehensive institutional system from three aspects: algorithm interpretability, data governance, and human-machine boundaries, ensuring that technology is merely a protector of fairness, not a culprit. Only by maintaining a careful balance between technological rationality, legal ethics, and manufacturing standards can digital justice and automated quality control become powerful driving forces for social fairness and industrial precision. This equilibrium ensures that explainable algorithmic decision-making serves as a reliable foundation for both judicial integrity and the high-quality production of spinning and weaving mills. A "black box" system cannot, by definition, support judicial integrity; only when the decision-making logic is traceable and subject to human oversight can it be considered a "reliable foundation" for digital justice.

AUTHOR CONTRIBUTIONS

Zeng Qinglin designed, collected and analyzed the data, and drafted the manuscript. Zeng Qinglin conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Zeng Qinglin participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] J. Zhou, L. Yang, Z. Liu, et al., "Research on Artificial Intelligence Assisted Judges' Decision-Making—Based on the Perspective of Sentencing Deviation Identification," Journal of Economic Management, vol. 2, no. 2, pp. 197-218, 2023.
[2] X. Zhang and S. Feng, "The dilemma and breakthrough of algorithmic justice: human-machine collaborative governance of generative AI in public libraries," Information Theory and Practice, vol. 48, no. 12, pp. 34-41, 2025.
[3] L. Li and X. Zhang, "Algorithmic injustice: problem types and governance paths," Journal of Dialectics of Nature, vol. 47, no. 5, pp. 76-83, 2025, doi: 10.15994/j.1000-0763.2025.05.009.
[4] Y. Peng, C. Yang, and G. Yu, "Exploration of accountability system for clinical auxiliary decision support system based on artificial intelligence," Journal of Naval Medical University, vol. 46, no. 8, pp. 963-969, 2025.
[5] L. Niu, "Judicial Application and Jurisprudential Examination of Artificial Intelligence," Journal of Henan Normal University (Philosophy and Social Sciences Edition), vol. 50, no. 5, pp. 56-62, 2023.
[6] L. Lei, "Can judicial artificial intelligence achieve judicial fairness? Political and Legal Review," 2022; (4): 72-82.
[7] H. Yang, "Expected effectiveness, potential risks and development direction of large-scale model-assisted sentencing," Journal of Xidian University (Social Science Edition), vol. 35, no. 4, pp. 114-124, 2025.
[8] Z. Liu, "On the negative impact of algorithmic cognitive bias on the legal regulation of artificial intelligence and its correction," Politics and Law, vol. (11), pp. 50-65, 2022.
[9] K. Wang and Y. Xu, "Ethical challenges and solutions of medical artificial intelligence technology based on intelligent algorithms," Chinese Journal of Medical Ethics, vol. 38, no. 9, pp. 1127-1132, 2025.
[10] B. Fan and J. Li, "How can algorithmic transparency enhance public trust in the government? Public Administration Review," 2024; 17(1):4-24.
[11] J. Zhang, "The structural dilemma and systemic positioning of intelligent justice," Academic Exchange, vol. (7), pp. 103-111, 2020.
[12] K. Gao, "Functions, risks and improvement of judicial intelligence," Journal of Xi'an Jiaotong University (Social Sciences Edition), vol. 40, no. 6, pp. 145-152, 2020.
[13] X. Lin, "A study on judicial credibility in the context of artificial intelligence intervention : an investigation based on cognitive experiments," Political and Legal Forum, vol. 44, no. 1, pp. 81-93, 2026.
[14] J. Fang and T. Gao, "Algorithmic risks and governance paths of generative artificial intelligence-assisted administrative decision-making," Huxiang Forum, vol. 37, no. 1, pp. 99-111, 2024.
[15] L. You, "On the meaning, content and rule of law path of the centralization of judicial auxiliary affairs," Northern Jurisprudence, vol. 13, no. 4, pp. 114-128, 2019.