

Research on the Procedural Justification Regulation of Sentencing Recommendations under the Human-Machine Collaboration Model

Jie Xia

School of Law, Anhui University of Finance & Economics, Bengbu Anhui 233041, China.

Abstract

The application of artificial intelligence (AI) technology in prosecutorial case handling is deepening, and sentencing recommendation assistance has become an important part of smart prosecutorial construction. AI provides data support and reference for prosecutors to propose sentencing recommendations through detailed analysis of rules such as elements of a crime and sentencing circumstances, promoting the transformation of sentencing activities from experience-driven to data- and experience-driven collaborative processes. However, the deep involvement of intelligent sentencing assistance has also raised issues such as algorithmic black boxes, leading to opaque decision-making processes, algorithmic bias, and potential procedural legitimacy risks such as systemic injustice, unclear allocation of human-machine decision-making weights, and the erosion of parties' procedural rights. This paper uses the principle of procedural legitimacy as an analytical framework, exploring the procedural regulation path of sentencing recommendations under the human-machine collaborative model from three dimensions: ensuring the substantive procedural participation of parties in the results of algorithm-assisted sentencing, constructing an explainable and auditable mechanism for algorithm-assisted sentencing recommendations, and clarifying the allocation of judicial responsibility for human-machine collaborative sentencing recommendations.

Keywords

Human-machine collaboration, Sentencing recommendations, Procedural legitimacy, Algorithm black box, Criminal proceedings.

1. Introduction of the Problem

Sentencing recommendations are procedural opinions submitted by the People's Procuratorate to the court regarding sentencing of criminal suspects and defendants during criminal proceedings. Article 275 of the "Rules of Criminal Procedure of the People's Procuratorate" stipulates that if a criminal suspect pleads guilty and accepts punishment, the People's Procuratorate shall submit sentencing recommendations regarding the principal punishment, supplementary punishment, and whether probation should be applied, and these recommendations should generally be for a fixed sentence. The accuracy of sentencing recommendations is directly related to the effective operation of the leniency system for those who plead guilty and accept punishment and the realization of judicial credibility.

In recent years, artificial intelligence (AI) technology has had a wide-ranging and profound impact on prosecutorial work. From legal research and document generation assistance to the embedding of generative large language models into core aspects of case handling such as evidence review, sentencing recommendations, and screening of supervisory leads, AI is continuously driving the transformation of prosecutorial work from informatization to intelligence. The assistance provided in sentencing recommendations is a typical example of

this transformation. Intelligent auxiliary systems, through detailed analysis of rules such as elements of a crime and sentencing circumstances, establish sentencing influencing factors and reference standards for sentencing ranges, providing intellectual support for prosecutors in making sentencing recommendations. In this sense, sentencing activities are undergoing a paradigm shift from experience-driven to a data - experience collaborative approach.

However, technological empowerment is not value-neutral. While improving sentencing efficiency and uniformity, human-machine collaboration also raises deep challenges in terms of procedural legitimacy. Some scholars have pointed out that artificial intelligence in the judiciary faces institutional problems such as rigid rules, unclear human-machine decision-making weights, and black-box decision-making by algorithms ^[1]. In human-machine collaborative trials, judicial responsibility determination faces dilemmas such as vague subjective responsibility determination, unclear distinction of responsible subjects, and increased risk of shirking responsibility ^[2]. At the same time, the deep application of judicial big models may also lead to the mechanistic erosion of the rights of the parties. This erosion is different from the occasional rights damage caused by individual case handling oversights in traditional justice, but is a systemic risk formed by the technical characteristics of automated operation of big models and black-box operation of algorithms ^[3]. The above studies have revealed the procedural dilemmas of artificial intelligence in the judiciary in a general sense, but systematic research that focuses specifically on sentencing recommendations in the review and prosecution stage and uses procedural legitimacy as an analytical framework still needs to be deepened. This article uses the principle of procedural legitimacy as an analytical tool to focus on the formation process of sentencing recommendations during the review and prosecution stage, examines the procedural legitimacy risks faced by human-machine collaborative sentencing recommendations, and explores corresponding procedural regulatory paths.

2. The Practical Logic and Procedural Tension of Human-Machine Collaborative Sentencing Recommendations

2.1. Practical Forms of Intelligent Sentencing Assistance

The technical principle of the intelligent sentencing assistance system can be summarized as "element extraction—factor quantification—weight assignment—reference output". The system comprehensively captures different sentencing reference factors based on the facts of the case, refines them into different sentencing weights, and uses a computational model to initially derive the sentence and type of punishment. In this process, the system establishes sentencing influencing factors and range reference standards through elemental analysis of rules such as constituent elements and sentencing circumstances, promoting a shift in sentencing from "same sentence for the same case" in similar cases to "same sentence for the same circumstances" in cases with similar circumstances.

Currently, the role of artificial intelligence (AI) in sentencing recommendations is explicitly defined as an "auxiliary tool." Practitioners point out that AI is positioned as a "digital assistant," not a "decision-making body." In case-handling stages such as the final determination of sentencing recommendations, judicial involvement must be strengthened, requiring prosecutors to personally conduct interrogations, review case files, and verify key evidence. Simultaneously, differentiated application rules need to be established based on the characteristics of different stages of prosecution, limiting the weighting of algorithmic models at the sentencing recommendation stage to ensure that technology always remains in an auxiliary role. However, a tension exists between the institutional declaration of its auxiliary role and the depth of technological intervention in practice, warranting careful examination.

2.2. The Four Tensions Facing Procedural Legitimacy

Procedural legitimacy, as a core principle of criminal proceedings, requires participatory, informed, reviewable, and accountable decision-making processes. The practice of human-machine collaborative sentencing recommendations faces structural tensions in all four dimensions mentioned above.

First, the tension between the algorithm black box and the transparency of the procedure. The algorithm black box refers to the opacity of the decision-making logic inside the algorithm and the inexplicability of the reasoning and argumentation process. At present, judicial artificial intelligence relies on closed proprietary algorithms. In addition to the program developers, judicial personnel also find it difficult to fully understand its operating logic. In the sentencing recommendation scenario, how the algorithm determines the selection criteria for sentencing factors, how it allocates weights among factors, and how it calculates the reference sentence are all key links that constitute cognitive barriers for prosecutors and parties involved. This opacity directly erodes the requirements of procedural legitimacy for the understandability and reviewability of the decision-making process [7].

Secondly, there is the tension between algorithmic bias and procedural fairness. Algorithmic bias refers to the systemic injustices that artificial intelligence systems produce towards specific groups during data processing and decision-making. In sentencing recommendations, if historical case data itself contains structural biases such as regional sentencing differences and inconsistent enforcement standards, the algorithm model trained on this basis will solidify or even amplify these existing biases. Furthermore, the design of the algorithm model involves the selection and weighting of case characteristics, a process imbued with the designer's value presuppositions. If diverse perspectives are not adequately considered, it may lead to systemic disadvantageous treatment of specific groups. Such implicit biases are difficult to detect and correct through traditional procedural mechanisms, yet they can be repeatedly reproduced through each sentencing recommendation.

Third, there is a tension between the unclear allocation of human-machine decision-making weights and the exclusivity of judicial judgment. When artificial intelligence provides reference suggestions on sentencing issues, must prosecutors refer to them? When prosecutors' judgments are inconsistent with the suggestions of artificial intelligence, do they bear an excessive obligation to explain? [2] The above questions lack clear answers within the current regulatory framework. In practice, if the sentencing recommendations deviate from the reference range of the intelligent auxiliary system, prosecutors may face pressure for case handling assessments or accountability risks; if the system's suggestions are fully adopted, the prosecutor's independent judicial judgment right is essentially undermined. This ambiguous state of power and responsibility allocation may cause the "assistance" positioning to be distorted into "domination" in practice.

Fourth, there is the tension between the parties' procedural rights and the algorithmic decision-making mechanism. During the formation of sentencing recommendations, criminal suspects and their defense lawyers are legally entitled to participate in sentencing negotiations and express their opinions. However, when sentencing recommendations are partly generated by algorithms, the parties face a complex decision-making structure of "algorithmic prediction + prosecutorial confirmation," making it difficult to understand the algorithm's decision-making logic and to substantially challenge the algorithm's output. This erosion is not an occasional individual violation of rights, but a systemic risk embedded in the operational logic of the technological system, characterized by its persistence, universality, and insidious nature.

3. Theoretical Basis for the Regulation of Human-Machine Collaborative Sentencing Recommendation Procedures

Procedural legitimacy provides a normative basis for regulating human-machine collaborative sentencing recommendations. In the formation of sentencing recommendations, procedural legitimacy includes at least four basic elements: First, participation—participants with a stake in the sentencing outcome have the right to participate in the formation process of the sentencing recommendation and to present arguments and evidence favorable to their side; second, the right to know—participants have the right to know the basis and reasons for decisions affecting their rights; third, the right to object—participants have the right to object to the sentencing recommendation and request a review; and fourth, accountability—when a sentencing recommendation is erroneous, the responsible party can be traced and procedural remedies can be initiated.

The above requirements are guaranteed in the traditional judicial context through litigation procedures and judicial ethics. However, the human-machine collaboration model has changed the cognitive structure of judicial decision-making. Some scholars have pointed out from an epistemological perspective that the datafication of cognitive sources, the algorithmization of cognitive processes, and the probabilization of cognitive results in large models have a structural tension with the diversity of cognitive sources, the procedural nature of cognitive processes, and the certainty of cognitive results of judicial personnel [4]. Specifically, in the sentencing recommendation scenario, this tension presents as a triple misalignment: the probabilistic output of the algorithm is difficult to replace the value judgment of the prosecutor based on the whole case; the datafication of the algorithm is difficult to cover the incommensurable rational elements in individual cases; and the black box decision of the algorithm is difficult to meet the normative expectations of procedural legitimacy for transparency and reviewability.

In response to the aforementioned tensions, the academic community has developed two main regulatory approaches. The first is the "auxiliary role theory," which emphasizes preventing technological risks by clearly defining the boundaries of human-machine division of labor and argues that artificial intelligence should not replace the independent judgment of judicial personnel. The second is the "algorithmic due process theory," which advocates incorporating algorithmic decision-making into the normative framework of due process and absorbing technological risks through the modification of procedural mechanisms. Both approaches affirm the auxiliary role of artificial intelligence in sentencing, but the former focuses on passively avoiding risks by defining the functional boundaries of artificial intelligence, while the latter focuses on actively absorbing risks by reconstructing procedural mechanisms. Some scholars have further proposed that AI-assisted sentencing should be included in the scope of judicial review and "reformed through evidence"—based on the existing criminal procedure framework, positioning AI-assisted sentencing recommendations as specialized reports, and systematically regulating them from dimensions such as evidence rules, the subject of evidence presentation, the stage of evidence formation, cross-examination and authentication, and judicial liability.

This article argues that the procedural regulation of human-machine collaborative sentencing recommendations should follow three basic principles: the principle of auxiliaryity, where artificial intelligence can only be an auxiliary tool and the final decision-making power for sentencing recommendations belongs to the prosecutor; the principle of transparency, where the operational logic and decision-making basis of the algorithm should be explainable and auditable; and the principle of participation, where the procedural rights of the parties involved should not be weakened by technological intervention, but rather their substantive

participation in the human-machine collaborative decision-making process should be ensured through institutional design.

4. The Implementation of the Human-Machine Collaborative Sentencing Recommendation Procedure

4.1. Ensuring the parties' procedural participation in algorithm-assisted results

In the traditional sentencing recommendation procedure, the "Rules of Criminal Procedure of the People's Procuratorate" clearly require that before making a sentencing recommendation, the People's Procuratorate shall, in accordance with the law, solicit the opinions of the criminal suspect, defense counsel or duty lawyer, victim and their litigation agent, and inform them of the facts of the crime to be identified, the suspected crime, the sentencing circumstances and corresponding mitigating and aggravating measures, the proposed sentencing recommendation and its legal basis. In the human-machine collaborative model, the above procedural rights should extend to participation in and questioning of algorithm-assisted results.

First, the obligation to inform parties of algorithm-assisted results. When sentencing recommendations are formed using intelligent assistance systems, the prosecution must inform the parties involved of this fact and explain the system's role and impact in the case. When using AI-assisted systems during the trial process, the overall operational mechanism of the AI-assisted judicial system must first be explained to the parties involved to ensure their right to know and their right to choose. This requirement also applies to the sentencing recommendation procedure during the review and prosecution stage.

Second, the right of parties to object to algorithmic results. Parties should be granted the procedural right to object to the output of intelligent assistance systems. If a party believes that the algorithm-assisted result is biased or unfair, they have the right to request the prosecutor to explain the basis for the result and to submit rebuttal evidence or opinions. Simultaneously, if the defendant requests it, an expert assistant should be provided, and the defendant's right to object should be guaranteed. In sentencing recommendation scenarios, this right specifically includes: the party's right to understand the sentencing factors and their weighting based on the algorithm; the right to request supplementary explanations for individual case circumstances not covered by the algorithm; and the right to request the prosecutor to explain the reasons for adopting or deviating from the algorithm's recommendations.

Third, procedural safeguards for the verification of algorithm-assisted sentencing. A mechanism for verifying AI-assisted judicial decisions should be established to ensure that defendants and their lawyers have the right to request review, questioning, and cross-expert consultation. While the sentencing recommendation stage does not involve courtroom verification, a similar procedural safeguard mechanism can be established—during sentencing negotiations, the parties involved and their lawyers have the right to question the relevant outputs of the AI-assisted system, and the prosecution should respond.

4.2. Constructing an interpretable and auditable mechanism for sentencing recommendation algorithms

The black box of algorithms is the core obstacle restricting the legitimacy of human-machine collaborative sentencing recommendation procedures. Solving this problem can be achieved by establishing institutional mechanisms at the levels of interpretability and auditability.

At the level of interpretability, intelligent auxiliary systems involved in sentencing recommendations should be required to possess basic interpretability functions. The procuratorate should improve the interpretability of algorithms to overcome the "black box"

dilemma and build public trust in the judiciary. Simultaneously, artificial intelligence systems involved in judicial decisions should possess basic interpretability functions, enabling them to clearly demonstrate their operational mechanisms, reasoning structures, decision weights, and the logical basis influencing the generation of conclusions. Specifically, in sentencing recommendation scenarios, intelligent auxiliary systems should be able to explain: what are the main factors influencing sentencing, how are the weights of each factor determined, how is the reference sentence calculated, and the degree to which different circumstances affect the sentence.

At the auditability level, a full-process record-keeping and auditing system for algorithm-assisted sentencing should be established. The algorithm system should be required to retain complete decision-making support logs, including input data versions, model versions, multiple possible output options and confidence levels, etc., to provide technical possibilities for post-event professional auditing and accountability [5]. In the sentencing recommendation scenario, this means that every operation of the intelligent assistance system should leave a traceable record: the input case elements, the reference suggestions output by the system, and the prosecutor's adoption or deviation from the suggestions. These records are not only the basis for post-event supervision and accountability, but also the basis for the parties to exercise their right to know and right to object.

At the level of algorithm supervision, a unified algorithm technology standard and supervision system should be established. The state should formulate a unified technical standard for generative artificial intelligence technology in the field of criminal trial, give full play to the top-level design role of the Supreme People's Court in the construction of smart justice, and integrate "algorithm transparency" and "explainability" into the unified technical standard. [8] This suggestion also applies to the sentencing intelligent auxiliary system of the procuratorial system. A unified algorithm standard and supervision system can prevent algorithm bias and algorithm black box risks from the source, and ensure that sentencing recommendations in different regions and different cases have consistent explainability and auditability at the algorithm level.

4.3. Clarify the allocation of judicial responsibility for sentencing recommendations based on human-machine collaboration

The core principle of the judicial accountability system lies in "letting the adjudicator decide and the decider be responsible." The "Opinions of the Supreme People's Court on Regulating and Strengthening the Judicial Application of Artificial Intelligence" clearly states that "judicial responsibility ultimately rests with the decider." In sentencing recommendation scenarios, the final decision-making power for sentencing recommendations belongs to the prosecutor, and judicial responsibility should also be borne by the prosecutor. However, the intervention of artificial intelligence complicates the determination of responsibility—technical factors such as algorithmic black boxes and data biases make it difficult to trace the causes of wrongful convictions, and the standards for judging whether prosecutors fulfilled their duty of care and whether gross negligence occurred become blurred. Current regulations "only establish the ultimate attribution of responsibility in principle, without specifying the specific standards and scope of responsibility, resulting in a certain degree of ambiguity in the accountability process." In response, the responsibility allocation mechanism can be improved from the following aspects:

First, a distinction must be made between algorithmic design flaws and prosecutorial oversight negligence. If the error in sentencing recommendations stems from algorithmic design flaws or training data issues within the intelligent assistance system itself, the responsible party should be the system's designer, developer, or deployer. If the error arises from the prosecutor's failure to fulfill reasonable review obligations, such as failing to detect obvious errors in the

algorithm's output or failing to consider significant sentencing factors not covered by the algorithm, then the responsibility should lie with the prosecutor. Simultaneously, human-machine responsibility should be differentiated based on product liability, with the producers, sellers, designers, managers, maintainers, and data providers of artificial intelligence bearing corresponding product liability.

Second, establish the prosecutor's reasonable duty of care and review standards. When prosecutors use intelligent auxiliary systems to make sentencing recommendations, they should fulfill their reasonable duty of care, including verifying whether the case elements input by the system are accurate, reviewing whether the reference recommendations output by the system are consistent with the facts of the case, and considering the special circumstances of the case that the system may not cover. The judgment of whether the prosecutor has fulfilled his reasonable duty of care should be based on whether "a prosecutor with general professional ability can find the error in the same situation". If the prosecutor has fulfilled his reasonable review duty but still cannot find the error, he can be exempted from liability; if the error can be found with general professional judgment but is still accepted, then the prosecutor has committed gross negligence [2].

Third, establish a basis for accountability based on full-process traceability. The operation of intelligent auxiliary systems should be fully traceable, including input data, running time, output results, prosecutor's review opinions and final decisions. These records are not only the basis for judging whether prosecutors have fulfilled their reasonable duty of care, but also the institutional basis for tracing algorithm problems. Technically, ensuring full-process traceability can alleviate the situation where there is no way to hold anyone accountable [6].

Fourth, technology vendors should be included in the accountability system. When errors in sentencing recommendations stem from technical flaws in intelligent assistance systems, holding only prosecutors accountable is neither fair nor reasonable. Technology vendors should be included in the accountability framework, establishing a responsibility-sharing mechanism among algorithm designers, deployers, and users. A tiered, dynamic accountability system should be established, setting "product liability" for algorithm providers. Referring to product liability principles, they should be required to ensure the basic security, fairness, and reliability of the algorithm in the intended judicial application scenarios.

5. Conclusion

The regulation of procedural legitimacy of sentencing recommendations under the human-machine collaboration model is an important issue facing criminal justice in the digital age. While improving judicial efficiency and promoting sentencing uniformity, intelligent sentencing assistance has also triggered procedural legitimacy risks such as algorithm black box, algorithm bias, unclear human-machine decision weight, and erosion of the parties' procedural rights. These risks fundamentally challenge the core values of procedural legitimacy, such as participation, awareness, auditability, and accountability. To address these challenges, it is necessary to build a procedural regulation system from multiple levels while adhering to the auxiliary positioning of artificial intelligence: ensuring the parties' procedural participation in the algorithm assistance results and ensuring that the right to know and the right to object are not weakened after the intervention of technology; building an interpretable and auditable mechanism for algorithm assistance to crack the algorithm black box; clarifying the allocation of judicial responsibility for human-machine collaborative sentencing recommendations to ensure that responsibility is traceable and accountable. Only in this way can a dynamic balance be achieved between technological empowerment and procedural justice, so that artificial intelligence can truly become an aid to improving judicial fairness rather than a hidden danger that erodes procedural legitimacy. The key to the success of smart justice lies in designing a

human-machine interactive decision-making system that conforms to judicial principles ^[1]. This issue is of particular urgent practical significance in the crucial aspect of sentencing recommendations.

References

- [1] Ding, X. D. (2023). Smart justice under human-computer interaction decision-making. *Legal Science (Journal of Northwest University of Political Science and Law)*, 41(4), 58–68.
- [2] Chen, Z. J. (2026). Intelligent response of judicial responsibility in human-machine collaborative trial model. *Jiaotong University Law Journal*, (2), 34–45.
- [3] Yu, S. Q. (2025). Institutional reshaping of the judicial big model. *Qiushi Journal*, 52(5), 133–147.
- [4] Wang, C. G. (2025). Cognitive path and norms of judicial application of large model. *Chinese Journal of Criminal Law*, (5), 159–176.
- [5] Li, J. B. (2026). On the paradigm shift of legal interpretation of large models in digital justice. *Journal of Suzhou University (Philosophy and Social Sciences Edition)*, 47(2), 88–99.
- [6] Zhang, L. H. (2022). Concerns and countermeasures regarding technology dependence in smart justice. *Legal System and Social Development*, 28(4), 180–200.
- [7] Yi, J. L. (2025). On the construction of trust in human-machine collaborative trial. *Chinese Jurisprudence*, (2), 148–167.
- [8] Chen, H., Liu, C., & Qi, H. Y. (2024). Advantages, risks and regulation of AIGC technology in assisting criminal trials—taking ChatGPT as an example. *Journal of Yangtze University (Social Sciences Edition)*, 47(4), 113–118.