https://doi.org/10.37420/j.mlr.2026.020

Artificial Intelligence-Assisted Sentencing: Rationale, Procedural Justice Risks, and Mitigation Pathways

Gao Qinghua*
Minzu University of China, Beijing, China
***Corresponding author E-mail: gqh666999@163.com**

Abstract:
The development of smart justice and the continuing reform of sentencing standardization have accelerated the practical application of artificial intelligence in sentencing. AI-assisted sentencing operates by transforming legal rules into computable data elements and by learning adjudicative patterns from large volumes of judicial data. The explainability of such systems rests, in significant part, on the formal-rational transformation between legal rules and algorithmic data. Yet the introduction of artificial intelligence into sentencing also reshapes traditional judicial procedures and gives rise to a range of procedural justice concerns, particularly in relation to judicial agency, the legitimacy of decision-making authority, algorithmic opacity, and the reviewability of machine-generated recommendations. These risks require a clear institutional response. Artificial intelligence should remain confined to an auxiliary role in adjudication, while human judges must retain final decision-making authority and responsibility. At the same time, the quality of judicial data and the explainability of algorithmic processes should be improved. Procedural safeguards must also be strengthened by ensuring defendants' rights to information, participation, explanation, and meaningful challenge, together with effective systems of judicial review, independent oversight, and institutional feedback. These measures can help reconcile the efficiency gains of AI-assisted sentencing with the procedural requirements of legitimate criminal adjudication.

Keywords:
sentencing; smart justice; procedural justice in sentencing; artificial intelligence; AI-assisted sentencing

1 Introduction

1.1 Policy Background

The standardization of sentencing depends on the sharing of judicial data, the harmonization of sentencing standards, and greater accessibility of public data, while AI-assisted sentencing presupposes the broader integration of artificial intelligence into judicial practice. Against the policy backdrop of building a "digital rule of law" and promoting "smart justice," the Supreme People's Court, the Ministry of Justice, and other authorities have issued policy instruments supporting the development of judicial data platforms and data-sharing mechanisms. Technologies such as big data analytics, artificial intelligence, and cloud computing have conse-

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

quently become increasingly embedded in judicial practice, contributing to the digital and intelligent transformation of China's justice system.

Since the nationwide advancement of sentencing standardization reform in 2010, the establishment of the system of leniency for defendants who plead guilty and accept punishment has provided further impetus for reform. The growing use of big data and intelligent technologies in the formulation of sentencing recommendations has improved the consistency and precision of sentencing. In particular, AI-assisted sentencing systems can reduce the routine workload of judicial personnel while facilitating more efficient case processing and more standardized sentencing outcomes. As plea-based cases have increased and machine-learning techniques have continued to develop, the conditions for more stable and refined sentencing assistance have gradually taken shape. Against this background, AI-assisted sentencing has become an increasingly important component of smart justice.

1.2 AI-Assisted Adjudication in the Era of Weak AI

Supported by policies encouraging the use of artificial intelligence in sentencing, judicial authorities across China have developed a range of intelligent sentencing-assistance systems. At present, however, judicial AI remains within the stage of "weak AI" and has not acquired the capacity for genuinely autonomous legal decision-making. Such systems continue to depend on judicial data supplied, structured, and curated by humans for algorithmic training. Their principal functions therefore remain concentrated on information processing and decision support. At this stage, judicial AI applications can broadly be divided into two categories: legal information retrieval systems and legal expert systems.[1]

Expert systems represent one of the earliest practical applications of artificial intelligence. They may be understood as computer systems that integrate specialized knowledge and professional experience within a particular domain. In the legal context, expert systems process large volumes of judicial data and are refined through expert-guided training in order to reproduce, to a limited extent, the reasoning process from "known facts" to "legal conclusions." By entering relevant case information, users may obtain outputs such as similar-case recommendations and suggested sentencing ranges.

Chinese courts and procuratorates have accordingly developed various forms of AI-assisted case-handling systems. Representative examples include Shanghai's Criminal Case Intelligent Assistance System, commonly known as the "206 System," introduced in 2017, and the Beijing courts' "Rui Faguan" intelligent judicial analysis system. By undertaking repetitive and comparatively routine tasks, such systems allow judges and other judicial personnel to devote greater attention to the substantive and procedural issues of individual cases, thereby improving the overall efficiency of adjudication.

2 Theoretical Foundations: The Transformation Between Legal Rules and Algorithmic Data Through Formal Rationality

2.1 The Explainability of AI-Assisted Sentencing

The explainability of artificial intelligence in judicial decision-making refers to the extent to which the

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

reasoning processes underlying data-driven outputs are transparent, accessible, and intelligible to relevant human actors. Explainability has particular importance in the judicial context because legal decisions cannot ordinarily derive their legitimacy from outcomes alone. Courts are expected to provide reasons, relate factual findings to legal norms, and make visible the principal considerations supporting the exercise of public power. If artificial intelligence contributes to that process, its role must likewise be capable of being understood and evaluated.

Scholars have approached the concept of explainability from different perspectives. Some identify four central objects of explanation: the judicial data on which a system relies, the goals it is designed to pursue, the algorithmic model through which inputs are processed, and the cognitive relationship between judges and machine-generated recommendations.[2] Others describe artificial intelligence as operating through three principal components—expert knowledge, data, and algorithms—and argue that judicial application requires legal provisions, adjudicative rules, and standards of judgment to be classified, decomposed, and expressed in structured form.[3] Although these approaches emphasize different elements, they converge on a fundamental point: judicial AI cannot operate unless legal norms and adjudicative experience are translated into forms that computational systems can process.

The explainability of AI-assisted sentencing therefore has a structural foundation in the relationship between legal rules and algorithmic data. Legal rules are expressed in normative language, while algorithmic systems operate through formalized variables, computational relationships, and structured inputs. For artificial intelligence to assist sentencing, a bridge must be constructed between these two forms of representation. That bridge is made possible, at least in part, by the formal rationality shared by legal reasoning and algorithmic operations.

2.1.1 The Formal Rationality of Law

Both legal norms and legal reasoning display characteristics of formal rationality. Legal norms ordinarily specify conditions under which particular legal consequences follow. Although the elements of a legal rule may not always be expressed in an identical textual structure, legal application requires decision-makers to identify relevant conditions, determine whether factual circumstances satisfy those conditions, and derive legal consequences from that relationship. In criminal adjudication, this structure is particularly visible. The application of substantive criminal law requires the identification of legally defined offence elements, the establishment of relevant facts, and an assessment of whether those facts correspond to the normative requirements of the offence.

Chinese criminal-law scholarship has often described this process through the concept of correspondence between factual circumstances and legally defined constituent elements. Legal interpretation determines the content and scope of the relevant rule; fact-finding establishes what occurred in the particular case; and legal reasoning connects the two.[4] The process is not reducible to mechanical deduction, since interpretation and evaluation remain necessary, but it nevertheless possesses an identifiable structure. Legal language categorizes facts, facts are assessed against normative criteria, and legal consequences are justified through reasoned connections between the two. It is in this sense that legal reasoning possesses formal rationality.

This formal dimension of law enables legal norms to be represented, at least partially, in structured form.

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

Legally relevant concepts can be identified, divided into elements, and associated with specified consequences. Sentencing itself contains numerous features that lend themselves to such structuring. The nature of the offence, the amount of harm caused, the defendant's role, prior criminal history, confession, restitution, surrender, cooperation, and other mitigating or aggravating circumstances can be represented as variables. Sentencing guidelines and judicial practice may then associate those variables with starting points, adjustments, ranges, or other consequences. The resulting model remains an imperfect representation of human adjudication, but its existence demonstrates the extent to which aspects of sentencing are capable of formalization.

Algorithmic rules display a comparable, though not identical, form of rationality. AI-assisted sentencing systems use historical judicial data and predefined computational procedures to generate recommendations for new cases. Long before the development of contemporary machine learning, scholars had explored the possibility of representing legal reasoning through symbolic logic and of simplifying complex bodies of law through formal systems.[5] Later developments in logic programming further demonstrated that rules derived from statutory law and precedent could be combined in computational models designed to address legal problems.[6]

Modern AI systems do not necessarily operate through a simple legal syllogism. Machine-learning models may identify statistical relationships that do not correspond neatly to the linear structure of "major premise–minor premise–conclusion."[1] They may also absorb patterns reflecting judicial discretion, institutional practice, and variables that are not expressly stated in legal rules. Nevertheless, such systems still rely upon formalized computational procedures. Inputs are represented in machine-readable form; models process those inputs according to specified architectures and parameters; and outputs are generated through reproducible computational operations.

In this sense, algorithmic decision-making can be understood as a modelled reproduction of aspects of human reasoning and past human decisions. It does not replicate judicial thought in its entirety. Nor can it reproduce the full normative and experiential dimensions of judging. Yet it converts selected features of past decisions into formal relationships capable of being applied to new cases. The algorithm itself operates through predefined or learned computational sequences and produces symbolic mappings between data inputs and outputs. This structured operation constitutes a form of formal rationality.

The relationship between legal and algorithmic rationality is therefore one of analogy rather than identity. Legal reasoning is normative, institutional, and justificatory; algorithmic reasoning is computational and model-based. Yet both depend upon structured relationships among categories, conditions, and consequences. It is this partial compatibility that makes the technical transformation of legal rules into algorithmic data possible.

2.1.2 The Technological Logic of AI Intervention in Adjudication

Because legal rules and algorithmic operations each possess formalizable structures, technological systems can translate selected legal elements into computational representations. Legal texts, factual allegations, evidentiary information, and sentencing circumstances can be processed through techniques such as natural language processing, knowledge graphs, entity recognition, and feature extraction. Through these methods, textual descriptions of legally relevant facts can be transformed into structured variables capable of entering

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

algorithmic models.

The construction of a case knowledge graph illustrates this process. Legal actors, conduct, consequences, procedural events, and relationships among them may be represented as structured entities. The system then identifies relevant features of a case and compares them with features extracted from prior cases. Large bodies of judicial decisions can provide training material from which models identify recurring associations between case characteristics, legal classifications, and sentencing outcomes. Where particular combinations of factual circumstances consistently correspond with particular sentencing ranges, the model can abstract those relationships into predictive rules or probabilistic associations.

This process can be represented conceptually as a movement from legal facts to structured variables and from structured variables to predicted outcomes. The relationship may be simplified as "X → Y": where a specified combination of legally relevant features constitutes X, a particular sentencing recommendation or probability distribution may constitute Y.[7] The apparent simplicity of that formulation should not obscure the complexity of the underlying model. Nevertheless, it illustrates the basic mechanism through which formalized legal language becomes interoperable with computational language.

The technological intervention of artificial intelligence in sentencing thus depends upon successive acts of translation. Legal language is first transformed into structured legal elements; those elements are converted into data; data are processed through algorithmic models; and algorithmic outputs are subsequently reintroduced into the legal process as recommendations. At each stage, choices must be made about what information is relevant, how it should be categorized, which relationships should be emphasized, and what type of output should be generated. These choices demonstrate why AI-assisted sentencing cannot be understood as purely technical. Its operation is based on a chain of legal and institutional judgments embedded within technical structures.

The shared formal rationality of law and algorithms therefore explains the technical possibility of AI-assisted sentencing, but it also reveals the source of its limitations. The computational system can process only those aspects of legal judgment that have been represented in forms accessible to it. Where legal evaluation depends upon context, moral judgment, institutional experience, or meanings that resist reliable formalization, algorithmic representation becomes more difficult. Explainability must therefore address not only how a model calculates an output, but also how legal concepts were transformed into the data on which that calculation depends.

2.2 The Content of Algorithmic Learning

2.2.1 Legal Language and Rules of Judicial Reasoning

The operation of AI-assisted sentencing requires more than the technical ability to process large volumes of data. It requires the system to engage, however indirectly, with the distinctive language and reasoning structures of law. Legal language has both procedural and argumentative characteristics. It functions within institutionally defined procedures, but it also seeks to justify outcomes through concepts recognized as legally relevant. Terms such as intention, negligence, principal offender, accomplice, surrender, confession, restitution, and dangerousness do not operate as ordinary descriptive labels. They carry legal meanings shaped by

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

statutes, doctrine, judicial interpretation, and procedural context.

Max Weber's well-known image of the "legal vending machine" provides an influential metaphor for highly formalized adjudication.[8] In this model, the facts of a case and the parties' claims are effectively inserted into a machine, and a legal judgment is produced in return. The image captures a longstanding aspiration of legal formalism: if rules are sufficiently clear and facts are sufficiently determinate, adjudication may appear capable of being reduced to structured application.

Contemporary artificial intelligence gives this metaphor renewed significance. A sentencing system may receive structured case information and generate a recommendation in a manner that superficially resembles Weber's machine. Yet the metaphor also exposes the danger of oversimplification. Judicial decision-making cannot be reduced to the mechanical processing of facts through predetermined rules. Even apparently straightforward cases require interpretation of legal concepts, assessment of evidence, evaluation of competing arguments, and justification of the relationship between facts and consequences.

Effective judicial AI must therefore do more than recognize words. It must distinguish legally significant information from irrelevant detail, relate factual descriptions to legal categories, and process the relationships among multiple legal factors. Natural language processing may assist in identifying entities and extracting facts, but legal reasoning requires an understanding of the normative significance assigned to those facts. For example, the fact that a defendant made a statement to the authorities may have very different sentencing consequences depending on whether the conduct legally constitutes voluntary surrender, confession, cooperation, or merely an ordinary admission.

The same difficulty applies to judicial reasoning. Courts do not simply apply isolated rules; they interpret rules in relation to facts, precedent, institutional standards, and competing considerations. AI-assisted sentencing therefore requires models capable of reflecting, at least in a limited way, the structures through which judicial decisions are made. This may include deductive reasoning, analogical comparison with similar cases, classification of legally relevant circumstances, and the identification of patterns in prior sentencing practice.

The role of legal expertise remains indispensable. Technical capacity alone cannot determine how offence elements should be defined, which circumstances are legally relevant, or how conflicts between rules should be resolved. Human legal knowledge must inform the construction, training, evaluation, and correction of AI systems. Judicial automation is thus not a process in which technology replaces legal reasoning. Rather, selected forms of legal reasoning are translated into technical systems and later reintroduced into adjudication as decision support.

2.2.2 Experiential Knowledge

The operation of AI-assisted sentencing also rests on a second assumption: judicial decision-making possesses a meaningful degree of predictability. If every sentencing decision were entirely unique and depended exclusively on irreducible individual judgment, historical data would have limited predictive value. AI-assisted sentencing becomes possible because judicial practice contains recurring patterns. Similar offences, comparable harm, similar offender characteristics, and recurring mitigating or aggravating circumstances tend to generate identifiable ranges of sentencing outcomes.

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

This does not mean that judicial decisions are perfectly predictable. Sentencing is influenced by the experience, preferences, and evaluative judgments of individual decision-makers, as well as by institutional practices and broader social contexts. Certain value-laden considerations cannot be reconstructed through formal syllogistic reasoning. Public controversy, perceptions of moral blameworthiness, changing social expectations, and judicial understandings of fairness may affect the treatment of particular cases in ways that resist precise codification.

Nevertheless, unpredictability is not absolute. Even where value judgments influence judicial decisions, their effects may produce observable patterns across large bodies of cases. Big-data analysis may reveal circumstances in which formal rules are applied more flexibly, factors that consistently affect sentencing severity, or recurring departures from nominally uniform standards. In this sense, algorithmic learning can extend beyond written legal rules and incorporate the experiential dimension of judicial practice.

Legal realism provides an important theoretical foundation for understanding this process. Llewellyn emphasized that legal rules can assist lawyers in investigating and predicting how judges are likely to act.[9] More broadly, legal realism challenged the idea that law exists only in authoritative texts. It directed attention to the conduct of legal officials and to the practical operation of law in real institutions. From this perspective, the meaning of law cannot be fully captured by studying rules in isolation from the way those rules are interpreted and applied.

This insight has direct relevance to AI-assisted sentencing. An intelligent system trained only on statutory provisions would capture the "law in books" but would remain unable to reproduce the full range of practices through which sentencing occurs. Judicial decisions, representative cases, sentencing recommendations, and other forms of adjudicative experience provide additional information about how legal rules operate in practice. By learning from such materials, AI systems attempt to identify "rules in action": regularities that emerge not solely from the text of law but from repeated institutional behaviour.

The incorporation of experiential knowledge also explains why AI-assisted sentencing is simultaneously attractive and problematic. On the one hand, historical data may improve prediction and help identify how similar cases have actually been treated. On the other hand, historical data may contain inconsistent, outdated, or biased practices. An algorithm that learns from past decisions may reproduce existing disparities as easily as it reproduces legitimate regularities. The use of experiential knowledge therefore requires careful attention to the quality, representativeness, and normative acceptability of the data from which the system learns.

AI-assisted sentencing ultimately operates through a combination of formal legal knowledge and experiential judicial knowledge. Legal rules establish normative boundaries, while judicial data reveal patterns of practical application. The ability to integrate these two sources explains the apparent capacity of AI to produce recommendations resembling human sentencing decisions. Yet the same combination also creates procedural risks. When the influence of past practice is embedded within complex algorithms, it may become difficult to determine why a particular recommendation was produced, whose judgment it reflects, and whether the recommendation remains consistent with current law and principles of fairness. These concerns lead directly to the procedural justice risks examined in the following section.

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

3 Procedural Justice Risks of AI-Assisted Sentencing

The increasing use of artificial intelligence in judicial decision-making has benefited from strong policy support, and the use of sentencing-assistance systems has undoubtedly contributed to improvements in judicial efficiency. Yet technological development does not itself establish procedural legitimacy. Legal expert systems capable of supporting judicial decisions have existed in various forms since the late twentieth century, but earlier systems relied heavily on expert-designed rules and human intuition. Contemporary machine-learning systems are more sophisticated, but their role in the legal field remains primarily auxiliary. Automated case management, information processing, similar-case retrieval, and sentencing support continue to dominate practice.

This gap between practical usefulness and adjudicative legitimacy is significant. AI-assisted sentencing may perform well as a technical instrument while still raising fundamental questions about the structure of judicial decision-making. Before artificial intelligence can reproduce the depth of legal knowledge, institutional understanding, and normative reasoning associated with human adjudication, its intervention in sentencing must be examined from the perspective of procedural justice. The most important concerns relate to judicial agency, the legitimacy of decision-making authority, responsibility for error, and the opacity of algorithmic recommendations.

3.1 The Paradox of Judicial Subjectivity

The first issue concerns the legitimacy of artificial intelligence as a participant in judicial decision-making. The technological functions of judicial AI have expanded considerably. Systems now contribute to the digitization of case information, intelligent case management, similar-case analysis, prediction of adjudicative outcomes, supervision of judicial practice, and the standardization of evidentiary or sentencing processes. Yet functional expansion does not mean that artificial intelligence has become an independent judicial subject.[10]

Judicial power is institutionally assigned to legally authorized human decision-makers and judicial bodies. An AI system does not possess legal personality, public authorization, professional responsibility, or an independent duty to interpret and apply the law. It follows that its role should remain auxiliary. The difficulty, however, lies in determining what "auxiliary" means once machine-generated recommendations become influential in everyday practice.

A sentencing recommendation produced by artificial intelligence may appear technically objective because it is generated through data and computation rather than personal intuition. That appearance can give the recommendation persuasive force. If an AI system identifies a particular sentencing range or predicts that comparable cases normally receive a particular punishment, a judge may feel institutional pressure to follow the recommendation. The question is therefore not merely whether AI formally possesses adjudicative authority, but how much practical weight its output acquires within the decision-making process.

The problem becomes especially visible when human and algorithmic judgments diverge. If an AI system recommends a substantially different sentence from that considered appropriate by the judge, should the judge be required to explain the departure? Should the algorithmic recommendation trigger additional review? Conversely, if a judge follows the recommendation, should reliance on the algorithm itself be explained? The

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

answers to these questions affect the location of actual decision-making authority. A formally non-binding recommendation may become functionally binding if judges perceive departure as burdensome, risky, or institutionally disfavoured.

A broader concern arises where courts and procuratorates employ standardized systems built on similar data and models. The use of common algorithmic standards may promote consistency, but it may also reduce the institutional diversity and independent judgment that form part of the criminal process. The constitutional and procedural structure of criminal justice relies upon different institutions performing distinct roles. If the same technological logic shapes investigation, prosecution, sentencing recommendation, and adjudication, the formal separation of institutional functions may become less meaningful in practice.

A second dimension of the problem is the risk of automation dependence. Judicial personnel work under significant caseload pressure. When an intelligent system offers rapid retrieval, case comparison, and sentencing recommendations, reliance on its output may become increasingly attractive. Over time, the convenience of automated recommendations may discourage independent legal reasoning. Judges may come to treat the system's result as a default position rather than as one source of information among many.

Such dependence risks weakening judicial agency. Courtroom argument may become less influential if algorithmic recommendations are treated as presumptively correct, while legal interpretation may become increasingly mechanical where decision-makers focus primarily on whether a case corresponds to predefined data categories rather than independently assessing whether those categories adequately reflect the legal and factual complexity of the individual case. Excessive reliance on algorithmic outputs may therefore reduce the space for judicial deliberation, legal interpretation, and meaningful engagement with the arguments presented by the parties. Nevertheless, this concern should not be overstated. AI-assisted systems can reduce unnecessary inconsistency and support judges in processing complex information; the real risk arises when assistance is transformed into deference. Procedural justice requires that the final decision remain the product of an accountable human decision-maker who has independently considered the facts, applicable law, and arguments of the parties. Preserving this requirement is essential to ensuring that technological assistance supports, rather than displaces, judicial judgment and responsibility.

3.2 Challenges to the Legitimacy of Decision-Making Authority

The second procedural concern relates to the legitimacy of authority exercised through AI-assisted decision-making. Judicial authority is not derived solely from the accuracy of outcomes. Courts possess authority because they are constituted by law, exercise publicly conferred powers, operate within defined procedures, and are accountable through institutional mechanisms. Individuals subject to judicial decisions are legally required to accept the authority of courts even when they disagree with the outcome.

Artificial intelligence does not possess an equivalent source of authority. An algorithm has no independent public mandate to determine an individual's rights, obligations, or liberty. Citizens are not subject to a legal duty to "obey artificial intelligence" as such. Even if an algorithm were statistically more consistent than some human decision-makers, technical performance would not by itself confer judicial power.

This distinction is especially important in sentencing because punishment represents one of the most coer-

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

cive exercises of state authority. The deprivation of liberty must remain attributable to institutions and individuals legally authorized to exercise penal power. If algorithmic recommendations were treated as independent decisions, the public character of adjudication would be obscured. The source of the decision would shift from identifiable legal institutions toward technical systems whose design, ownership, and operation may involve actors outside the formal judicial structure.

Questions of responsibility reinforce this concern. When an AI-assisted sentencing recommendation contributes to an erroneous or unjust outcome, the machine itself cannot bear legal or institutional responsibility. The system is the product of human design, data selection, programming, procurement, deployment, and use. Its operation reflects choices made by developers, administrators, prosecutors, judges, and other institutional actors. Under current conditions, responsibility must therefore remain with human actors rather than being transferred to artificial intelligence.

Control-based theories of responsibility support this conclusion. Responsibility ordinarily attaches to actors who possess sufficient knowledge of what they are doing and sufficient control over their conduct.[11] The users of AI-assisted sentencing systems decide whether and how to employ those systems, how to interpret their outputs, and whether to rely upon the resulting recommendations. They should therefore remain accountable for the consequences of such reliance.

The precise allocation of responsibility may vary according to the procedural stage. Where a procuratorial authority uses an AI system to formulate a sentencing recommendation, responsibility for the recommendation should remain with the prosecution. Where a court adopts or relies upon that recommendation, the adjudicative body must retain responsibility for the final sentence. Individual officials may likewise bear professional responsibility where they use automated outputs without adequate scrutiny. The key principle is that technological assistance must not create a responsibility gap.

The legitimacy problem is therefore twofold. First, artificial intelligence cannot independently acquire judicial authority merely by producing accurate or consistent predictions. Second, human institutions cannot rely on artificial intelligence in a manner that obscures who is responsible for the resulting exercise of public power. Procedural justice requires identifiable, legally authorized, and accountable decision-makers.

3.3 Data Security and Privacy Protection Risks

The application of artificial intelligence in judicial decision-making relies heavily on the collection, analysis, and processing of large volumes of data. Such data frequently contain sensitive information, including parties' personal information and detailed case materials, making data security and privacy protection matters of particular concern. Any unauthorized disclosure or misuse of such data may not only seriously infringe upon the privacy and reputational interests of the parties concerned, but may also undermine judicial fairness and public confidence in the justice system.

Although existing laws and regulations provide certain safeguards for data security and privacy protection, significant challenges remain in practice. For example, the collection and use of judicial data often lack sufficiently clear standards and procedural safeguards, while security measures governing data sharing and transmission may be inadequate, thereby increasing the risk of data breaches. Moreover, continuing technological

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

development has generated new forms of privacy and data-security risk. The capacity of deep-learning models to retain information may, for instance, create the possibility that sensitive data will be unintentionally disclosed, while the decentralized and immutable characteristics of blockchain technology may present additional challenges for the effective protection and management of personal information. Accordingly, as artificial intelligence becomes increasingly integrated into judicial decision-making, greater attention must be paid to data security and privacy protection. Relevant legal and regulatory frameworks should be strengthened and effectively implemented to ensure that judicial data are collected, processed, shared, and used in a lawful, secure, and appropriate manner.

3.4 Algorithmic Opacity and Barriers to Objective Accuracy

A third risk concerns the opacity of algorithmic decision-making. The capacities and limitations of self-learning systems cannot always be precisely identified, particularly where complex machine-learning models are involved. The internal process through which a model converts data into a recommendation may remain difficult to understand even for technically trained users. This problem is commonly described as the "algorithmic black box."

Opacity is especially problematic in judicial settings because a sentencing recommendation may influence an individual's liberty. A system may appear to generate an objective result, yet the result reflects decisions about training data, variable selection, model architecture, weighting, and evaluation. If those choices are invisible to judges and parties, the apparent neutrality of computation may conceal significant forms of human judgment.

Judicial AI systems require continuing testing, correction, and maintenance. Developers may modify algorithms to improve stability, accuracy, or efficiency. Yet the process of adjustment itself introduces human assumptions. Technical personnel inevitably make choices about what constitutes an error, which forms of disparity are acceptable, and which variables should be emphasized. Their professional experience and value judgments may therefore become indirectly embedded within the model.

Algorithmic bias may also be difficult to detect. If biased relationships exist in the training data, the system may reproduce them while presenting its outputs as data-driven conclusions. Because the model operates at scale, such distortions may affect large numbers of cases before they become visible. Correction is often retrospective: an error must first be detected before the system can be adjusted. The capacity of AI to standardize decision-making therefore creates the possibility that error itself may be standardized.

The problem is not limited to technical accuracy. From the perspective of procedural justice, opacity affects reviewability. If judges cannot understand the principal reasons for a recommendation, they cannot meaningfully evaluate whether reliance on it is justified. If defendants cannot identify the data or factors underlying the recommendation, they cannot effectively challenge its relevance or accuracy. If supervisory institutions cannot reconstruct the decision-making process, external review becomes less effective.

Algorithmic opacity therefore transforms a technical limitation into a procedural risk. The legitimacy of AI-assisted sentencing depends not only on whether the recommendation is statistically reliable, but also on whether its influence on the final decision can be examined, contested, and justified. This requirement pro-

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

vides the foundation for the mitigation pathways discussed below.

4. Mitigation Pathways: Establishing Legitimacy and Addressing Risks

4.1 Defining the Scope of Application: Balancing Efficiency and Justice

To address concerns over judicial subjectivity, the role of AI-assisted sentencing must be clearly confined to that of an adjudicative aid, while human judges must retain ultimate authority over sentencing decisions. In 2022, the Supreme People's Court issued the Opinions on Regulating and Strengthening the Judicial Application of Artificial Intelligence, which expressly established the principle of auxiliary use. It provides that artificial intelligence must not replace judges in adjudication and that AI-generated outputs may serve only as references for adjudication or judicial administration and supervision. Judicial decisions must continue to be made by judicial officers, adjudicative authority must remain vested in legally constituted judicial bodies, and ultimate responsibility must rest with the human decision-maker.

Accordingly, the role of AI in sentencing must be subject to substantive limits. Its comparative advantages are most apparent where consistency and standardization are particularly important, whereas individualized judgments should remain within the domain of judicial discretion. This distinction is especially relevant to the different stages of sentencing. At the initial stage, where the sentencing starting point and baseline sentence are determined, algorithmic modelling and data analysis may assist in promoting consistency across similar cases. By contrast, the assessment of discretionary mitigating or aggravating circumstances requires judges to draw upon legal expertise, practical experience, and contextual judgment. AI may therefore play a greater role in the more standardized stages of sentencing, while human judges must remain dominant where individualized evaluation is required. Such a division of functions offers a means of reconciling efficiency with substantive and procedural justice.

4.2 Improving Data Quality and Algorithmic Explainability

The opacity and limited explainability of algorithmic systems create the risk of an "algorithmic black box" in criminal proceedings. Such opacity cannot be entirely eliminated. The appropriate response therefore lies not only in attempting to "open" the black box, but also in explaining and constraining its operation.

Human judicial reasoning itself is not fully transparent. This is precisely why procedural law requires judges to provide reasons for their decisions, to ground judgments in facts and law, and to remain subject to institutional review. Similar requirements should apply to AI-assisted sentencing. The relationship between the information entered into an algorithmic system and the recommendation it produces should be sufficiently intelligible to permit meaningful scrutiny. AI-generated recommendations should also be capable of explanation in terms that allow judges and parties to understand the principal factors influencing the output.

High-quality judicial data are equally important. AI-assisted sentencing systems should be trained on reliable, representative, and sufficiently complete judicial databases. Existing data drawn from publicly available judgments or large judicial databases may be incomplete, omit legally significant information, or fail to reflect the full factual basis of adjudication. If algorithmic systems are trained on distorted or unrepresentative data, they may reproduce rather than correct existing inconsistencies or biases. Improving the quality and authen-

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

ticity of training data is therefore a necessary condition for increasing the reliability of AI-assisted sentencing and for ensuring that algorithmic recommendations reflect actual judicial standards.

4.3 Strengthening the Procedural Rights of Defendants

From the perspective of procedural law, the concept of "technological procedural justice" requires two complementary safeguards. The first is technological: databases, data authentication, human correction, and other technical mechanisms should be used to improve the transparency and reliability of algorithmic systems. The second is rights-based: the use of artificial intelligence must not diminish the effective participation of defendants in criminal proceedings. The protection of procedural rights must remain the normative foundation of AI-assisted sentencing.

The central concern is therefore the potential erosion of defendants' procedural rights once algorithmic systems are introduced into criminal justice. Technological innovation should not narrow the practical space within which defendants may participate, contest, and seek remedies. Legal safeguards must develop alongside technological change.

The European Union's 2018 General Data Protection Regulation provides a useful comparative reference by granting data subjects certain rights of control and choice in relation to algorithmic decision-making. Its underlying logic is that the exercise of technological power should remain constrained by individual rights. A similar approach should inform AI-assisted sentencing in criminal proceedings.

Defendants should therefore enjoy meaningful rights to be informed of the use of AI-assisted sentencing systems, to express their views on algorithmically generated recommendations, and to challenge the information or reasoning on which such recommendations are based. Where procuratorial authorities use AI systems to formulate sentencing recommendations, the data and principal analytical basis underlying those recommendations should be disclosed to the defence to the extent necessary for meaningful scrutiny.

Disclosure should not be reduced to the formal provision of large volumes of raw data. Where the prosecution possesses substantially greater technical capacity to interpret algorithmic outputs, procedural equality cannot be achieved merely by granting access to information that the defence is practically unable to analyse. Appropriate mechanisms should therefore enable the defence to understand the relevant data-processing methods and the principal logic of the algorithmic assessment. Only in this way can defendants exercise a substantive right to scrutinize AI-assisted sentencing rather than a merely formal right of access. Without such safeguards, existing inequalities between prosecution and defence may be aggravated by disparities in access to and capacity to analyse judicial data.

4.4 Strengthening Oversight and Feedback Mechanisms

A comprehensive and dynamic system of oversight and feedback is essential to maintaining fairness in AI-assisted sentencing. Multiple forms of supervision should therefore be incorporated into the design and operation of such systems, including internal review, independent third-party auditing, judicial scrutiny, and appropriate forms of public oversight.

Internal review mechanisms should regularly examine the transparency, legality, and reliability of algorith-

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

mic operations and identify potential errors or biases. Independent professional bodies should also be permitted to audit the processing of judicial data, the basic logic of algorithmic models, and the consistency of their outputs with applicable legal and ethical standards.

Judicial scrutiny is particularly important because the ultimate responsibility for sentencing remains with the courts. Courts should establish procedures for reviewing the use and adoption of AI-generated sentencing recommendations, especially where such recommendations appear unreasonable, insufficiently explained, or inconsistent with governing legal principles. Human decision-makers must be able to depart from algorithmic recommendations and should provide reasons where reliance on AI materially affects the sentencing outcome.

Effective feedback channels should also be available to defendants, lawyers, and other relevant participants. Their challenges and observations may reveal recurring errors, hidden biases, or practical deficiencies that would otherwise remain undetected. A combination of internal monitoring, independent auditing, judicial review, and user feedback can therefore provide an institutional framework within which AI-assisted sentencing systems remain subject to continuing legal and procedural control.

5 Conclusion

Within a framework of formal rationality, artificial intelligence offers important possibilities for improving the application of legal rules in criminal justice. Its capacity to process large volumes of judicial data, identify recurring patterns, and promote consistency across similar cases makes AI particularly attractive in the context of sentencing standardization. Nevertheless, existing technological limitations prevent AI-assisted sentencing from replacing human judges. Problems such as algorithmic opacity, bias, and excessive reliance on automated recommendations demonstrate that technological efficiency cannot displace the fundamental values of fairness, justice, and human judgment.

The appropriate response is therefore neither the unrestricted expansion nor the outright rejection of AI in sentencing. Its legitimacy must instead be constructed through institutional limits and procedural safeguards. AI should remain confined to an auxiliary role, while judges retain ultimate authority and responsibility for sentencing decisions. At the same time, improvements in judicial data quality and algorithmic explainability should be accompanied by meaningful protection of defendants' procedural rights and effective mechanisms of oversight and review.

The future development of AI-assisted sentencing thus requires a clear understanding of the respective strengths of human and algorithmic decision-making. Artificial intelligence may contribute efficiency, consistency, data-processing capacity, and pattern recognition, but human judges must remain responsible for legal interpretation, individualized evaluation, reason-giving, and the final exercise of judicial power. The proper objective is not to replace human adjudication with automated decision-making, but to integrate technological assistance into criminal justice without sacrificing the procedural foundations of legitimate sentencing.

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)

Funding

Graduate Research Projects of Minzu University of China: "Normative Interpretation and Systematic Construction of the Accused's Right to Meet from the Perspective of Mandatory Defense" (Project No. SZ-KY-X2025079).

References

[1]. Zhou, S.J., & Wu, X. (2019). The Possibilities and Limits of Artificial Intelligence in Judicial Decision-Making. Journal of East China University of Political Science and Law, 22(01), 53-66. (in Chinese)

[2]. Zhou, Y., & Zhang, X.J. (2023). The Dilemma of Explainability in Judicial Artificial Intelligence and Its Resolution. Law and Economy, (02), 3-20. (in Chinese)

[3]. Wu, T. (2020). Contextualism in Judicial Decision-Making in the Age of Artificial Intelligence—Also on the Explainability of Judicial Artificial Intelligence. Shanghai Legal Studies, (02), 102-127. (in Chinese)

[4]. Chen, X.L. (2023). Formal Rationality in Criminal Law Dogmatics. Peking University Law Journal, (02), 285-306. (in Chinese)

[5]. Brewer, S. (Ed.). (1998). Logic, Probability, and Presumptions in Legal Reasoning, London: Routledge.

[6]. Kowalski, R., & Sergot, M. (1990). The Use of Logical Models in Legal Problem Solving. Ratio Juris, 3, 201-218.

[7]. Wang, L.S. (2018). Technical Obstacles to the Development of Judicial Big Data and Artificial Intelligence. China Law Review, (02), 46-53. (in Chinese)

[8]. Coser, L.A. (1971). Masters of Sociological Thought: Ideas in Historical and Social Context, New York: Harcourt Brace Jovanovich.

[9]. Llewellyn, K.N. (1951). The Bramble Bush: On Our Law and Its Study, New York: Oceana Publications.

[10]. Zuo, W.M. (2018). Reflections on the Prospects for the Application of Legal Artificial Intelligence in China. Tsinghua University Law Journal, (02), 108-124. (in Chinese)

[11]. Fischer, J.M., & Ravizza, M. (1998). Responsibility and Control: A Theory of Moral Responsibility, Cambridge: Cambridge University Press.

© 2026 The Author(s). Published by Mason Publish Group. This open access work is licensed under a Creative Commons Attribution 4.0 International License. (http://creativecommons.org/licenses/by/4.0/)