Research on the legality and applicable limits of AI intervention in criminal justice

Yun Pei

Graduate School, EMILIO AGUINALDO COLLEGE, 006302, Manila, Philippines,
Email: 125354624@qq.com

Abstract. With the acceleration of globalization, transnational crimes are becoming increasingly complex and diverse, especially drug trafficking, cyber fraud, telecommunications fraud, money laundering and human trafficking, which are frequent and pose a serious threat to the security of the international community. As an important neighbor and regional cooperation partner of China, the Philippines is particularly prone to transnational crimes, and judicial assistance between China and the Philippines has become a key mechanism for dealing with cross-border crimes. Based on legal system analysis and case studies, this paper systematically sorts out the current status of judicial assistance between China and the Philippines, and deeply explores the main issues such as differences in legal systems, political and diplomatic factors, technical barriers and extradition disputes. It then proposes paths and countermeasures such as improving the docking of bilateral legal systems, building an information sharing and law enforcement cooperation platform, promoting the unification of technical standards for case handling, enhancing human rights protection and judicial mutual trust, and the strategic use of multilateral cooperation platforms. The study aims to provide theoretical support and practical reference for transnational judicial cooperation between China, the Philippines and related regional countries, and promote the modernization of regional governance systems and governance capabilities.

Keywords: AI; Criminal Justice; legality; Applicable Limits

1. Introduction

Since the 21st century, artificial intelligence (AI) technology has developed rapidly around the world. From the initial machine learning algorithms and speech recognition to today's deep neural networks, natural language processing and big data analysis, AI is penetrating into all areas of society at an unprecedented speed and breadth, especially in public governance and law enforcement systems. As the most serious area in the legal system and the one that is most concerned with the balance between state power and individual rights,

criminal justice is experiencing an unprecedented wave of technological innovation. At present, AI technology has been gradually introduced and piloted in many core links of criminal justice, such as investigation, sentencing, judgment, and crime prediction. Its technical performance and institutional significance are causing widespread attention and deep reflection.

In the investigation link, AI enables the police to efficiently and quickly lock and track suspects through technical means such as image recognition, behavior trajectory analysis, and face comparison. Intelligent police platforms represented by the "Snow Bright Project" and the "Eagle Eye System" have not only improved the efficiency of investigation, but also reshaped the detection model of criminal cases. In the sentencing and adjudication phase, several courts have piloted the construction of intelligent sentencing assistance systems, which assist judges in case comparison and sentencing reference by combining big data with historical case algorithm models. Even in the execution phase, AI is used for recidivism risk prediction and parole assessment, forming an intelligent application chain covering the entire criminal justice process.

However, while AI technology brings efficiency improvements and information integration advantages, it also poses an unprecedented impact on traditional criminal justice concepts, operating mechanisms and legal value systems. On the one hand, AI can achieve positive effects such as information symmetry, case handling standardization, and reduced judicial resource consumption to a certain extent; but on the other hand, its algorithmic logic is opaque, technical bias is difficult to eliminate, and human discretion is compressed, which inevitably raises concerns about major rule of law issues such as judicial justice, abuse of power, and procedural guarantees.

More noteworthy is that the rapid advancement of AI in the field of criminal justice often has the characteristics of "technology first, law lagging behind". This phenomenon of technology and system development being out of sync not only easily creates the real risk of "algorithms replacing judges" or "technology intervening in adjudication", but is also likely to damage the basic rights of the accused and weaken the public's trust in judicial justice in the absence of legal norms and procedural guarantees. These issues are directly related to the legitimacy and institutional limits of artificial intelligence's involvement in criminal justice.

In this context, this article intends to systematically sort out the real picture of AI's involvement in criminal justice with the dual dimensions of "legality" and "limits" as the core, deeply analyze the institutional dilemmas and legal conflicts it faces, clarify the constitutional

basis, legitimacy logic and legal authorization boundaries of artificial intelligence's involvement in the operation of judicial power, and on this basis, put forward operational institutional construction suggestions.

2. Literature Review

Research on artificial intelligence in the field of criminal justice has been an important topic in the legal and technical ethics circles in recent years. With the widespread pilot and promotion of AI technology in judicial practice, academic discussions on its legality, legitimacy and governance boundaries have become increasingly in-depth. This article systematically sorts out and reviews relevant research at home and abroad to clarify the content, methods and deficiencies of existing research results, and to clarify the theoretical position and innovation path of this study.

2.1 Current status of foreign research

In the international academic community, research on artificial intelligence in criminal justice started early, especially in Europe and the United States. Scholars represented by the United States began to pay attention to the applicability and legitimacy of AI in judicial procedures. As early as 2016, a research team at Princeton University in the United States conducted a critical study of the "COMPAS" risk assessment system used by Wisconsin, pointing out that the system has serious algorithmic bias in predicting the risk of recidivism, which may form a systematic discrimination against African-American defendants, triggering an in-depth discussion on "algorithmic justice" and "technical neutrality myth" [1].

The British and American academic circles generally reflect deeply on the problems of "dehumanization", "judicial mechanization" and "weakening of procedural justice" that may be caused by the involvement of AI in the judiciary from the perspective of legal philosophy and constitutional principles. Scholars such as Joshua Kroll proposed the "law of explainability", emphasizing that the defendant and the public should be protected from the right to know and cross-examination of the AI adjudication process [2]. Cary Coglianese and David Lehr further proposed the theoretical framework of "algorithmic governance" in their research, attempting to establish a set of algorithms use regulation systems applicable to governments and judicial institutions [3].

In the European Union, research on the judicial application of artificial intelligence mainly focuses on ethical regulation and human rights protection. In 2019, the European Commission

adopted the "Ethical Charter for Artificial Intelligence and the Judicial System", emphasizing that the judicial application of AI must follow the principles of fairness, transparency, controllability and accountability. In continental legal countries such as Germany and France, scholars mostly reflect on how AI technology can reconstruct the boundaries between judicial power and administrative power from the perspective of jurisprudence and the distribution of state power. For example, German jurist Günther Jakobs proposed the "responsibility vacancy" theory, questioning whether algorithms can bear the task of determining "subjective guilt" in criminal law, and emphasizing the irreplaceable nature of human judgment in the judicial process [4].

2.2 Current status of domestic research

In my country, with the construction of "smart courts", the pilot of "digital procuratorates" and the advancement of "smart policing" systems, the application scenarios of AI in the judicial system are constantly expanding, and corresponding theoretical research is also gradually unfolding. Domestic scholars mainly discuss the relevant issues of AI's involvement in criminal justice from three dimensions:

Focus on the empirical research of technology application paths and advantages and effectiveness. Some studies summarize the application value of AI in case screening, document generation, sentencing assistance, etc. through field investigations and case analysis. For example, some scholars pointed out that intelligent assistance systems can significantly improve judicial efficiency and reduce the problem of inconsistent judgments in similar cases; some studies also believe that AI can achieve rapid breakthroughs in major cases through image recognition and behavior analysis technology in criminal investigations, which has important social governance value.

Focus on the legal risks and regulatory issues of AI's involvement in justice. More and more scholars have begun to pay attention to the problems that AI may bring, such as weakening of procedural justice, unclear subjects of judicial responsibility, and insufficient technical transparency. For example, Professor Chen Ruihua proposed that the role of artificial intelligence as a "power agent" in criminal justice is still unclear [5]. If there is a lack of effective authorization mechanism and supervision system, it is easy to lead to the abuse of technology. Professor Wang Guisong emphasized that under the existing legal framework, AI's identity as a "quasi-judge" has natural legitimacy barriers, and its involvement in the entity judgment link should be strictly restricted [6].

Focus on the theoretical construction of algorithm ethics and data governance. Some scholars with interdisciplinary backgrounds in law and information science have attempted to construct a basic framework of "algorithmic justice" and "data rule of law". For example, some scholars advocate the introduction of "algorithmic explainability" as a standard for the judicial system to evaluate the legitimacy of technology, and propose that a special technical review mechanism should be established to guarantee the right to be informed and the right to cross-examination. At the same time, regarding the boundary issues of criminal data collection, storage and use, researchers have also called for the introduction of a "special law on criminal artificial intelligence" to fill the institutional gap.

In summary, although current research has achieved certain results in recent years, there are still some shortcomings, including insufficient systematicity and theoretical depth, the disconnect between legal regulation and technological reality, and the lack of discussion on constitutional foundations, power boundaries and human rights protection. This article aims to conduct a comprehensive and systematic academic discussion on the legitimacy basis and institutional limits of AI intervention in criminal justice, which not only responds to the challenges brought by technological development, but also serves the strategic needs of building a rule of law in China.

3. Results

3.1. Real-life application of AI in criminal justice

As a cutting-edge achievement of modern science and technology, artificial intelligence has begun to be deeply embedded in all aspects of my country's criminal justice. From investigation to trial, to execution and punishment prediction, AI technology is effectively changing the traditional judicial operation logic through algorithm models, data calculation, system automation and other methods. This chapter intends to start from three main stages, systematically analyze the real path and typical practices of AI intervention in criminal justice, and reveal the opportunities and potential risks of the rule of law it brings.

3.1.1 Application of AI in the investigation stage

In the investigation stage of criminal cases, AI technology has been widely used in case analysis, target tracking, evidence collection and other links, especially in sub-fields such as face recognition, voice recognition, and behavior prediction. It has shown strong technical

capabilities. AI provides unprecedented investigative means for public security organs through the rapid integration and deep learning of big data through algorithm models.

Deployment of face recognition and video tracking system. As a key intelligent security project led by the Ministry of Public Security, the "Snow Bright Project" realizes the trajectory monitoring and accurate identification of specific suspects by building a monitoring network covering urban and rural areas and supporting face recognition and identity matching systems. For example, in 2019, when the public security organs in a certain place in Sichuan Province cracked a major drug case, they used AI to compare the surveillance videos of the suspect entering and leaving the community in the past 6 months, and locked his activity trajectory in just 30 minutes, which greatly improved the efficiency of solving the case.

Integration of speech recognition and call analysis technology. In the fields of combating telecommunications fraud and cybercrime, the public security organs have gradually deployed intelligent voice analysis systems, which can transcribe and analyze the content of the suspect's phone calls in real time, thereby warning of potential high-risk behaviors. At the same time, such systems can also be used for voiceprint comparison of on-site recordings to provide auxiliary support for the chain of evidence.

Experimentation of behavioral prediction models. Some public security units have built "public opinion risk perception models" and "behavioral tendency scoring systems" based on big data platforms. By mining information such as social behavior, mobile trajectories, and online browsing habits of specific groups of people, possible abnormal behaviors or potential criminal tendencies can be predicted, thereby intervening in advance. Although such systems are currently mostly used in the field of public security, there is a trend of gradually expanding to criminal warnings.

3.1.2 AI's auxiliary function in the trial stage

Compared with the technical feature of "active intervention" in the investigation stage, AI mainly participates in judicial activities in the form of "assisted decision-making" in the trial stage. Its core lies in providing judges with similar case references, sentencing recommendations and intelligent document generation services through judicial big data and judgment models.

The promotion and application of "intelligent sentencing assistance system". Represented by the "similar case intelligent push system" led by the Supreme People's Court, the system can automatically push the sentencing standards and judgment logic of similar cases by inputting case information and legal provisions, providing reference for judges. For example, when

handling drunk driving cases, the Pudong New District Court of Shanghai compared hundreds of similar cases through the system, and the sentencing range output by the system was three to five months of detention. The judge made a judgment based on the case, which helped to improve the consistency and standardization of sentencing.

The construction of the "judicial big data platform". Many local courts have built judicial data centers, and through structured input of case materials, the integration of judges' case handling and automatic data analysis is realized. In the Hangzhou Internet Court, judges can retrieve similar cases, cited precedents, and relevant laws recommended by the system in real time during the trial, which improves the efficiency of judgment and the quality of the text.

Pilot attempts of the "intelligent trial platform". For example, the "intelligent trial system" developed by the Shanghai Higher People's Court and Alibaba Cloud can not only complete the classification and sorting of case materials and the comparison of legal application, but also quickly generate the first draft of the judgment document after the trial. The platform also has a certain natural language recognition capability, which can assist in identifying logical conflicts between evidence and assist judges in hearing complex economic crime cases.

3.1.3 AI attempts in the execution and prediction stage

In the link of penalty execution and recidivism risk control, AI technology has also begun to be introduced experimentally to improve the scientific nature of penalty decision-making and regulatory efficiency.

Community correction data modeling. Some regions have introduced AI behavior analysis systems in community correction work. By integrating the entry and exit information, psychological assessment reports, social behavior and other data of the corrected persons, a "behavior stability scoring model" is constructed to assist correction agencies in determining whether there is a risk of escaping supervision and re-offending. For example, the "intelligent community correction platform" piloted in a certain place in Guangdong establishes dynamic risk levels for key targets and adjusts the supervision frequency according to changes in the scores.

Exploration of the "artificial intelligence parole prediction system". Some detention centers and prisons have tried to introduce AI models to assess the possibility of recidivism of prisoners, and make parole recommendations based on their performance in prison, reform attitude, psychological tendencies and other factors. This mechanism aims to reduce subjective judgment factors and improve the fairness and security of the execution stage of punishment.

Widespread attention to the recidivism risk prediction system. Influenced by the US "COMPAS" model, some criminal policy research institutions in my country have also tried to develop a recidivism prediction system suitable for local areas to guide sentencing recommendations and post-sentence supervision. However, due to the complexity of data sources and the difficulty of publicizing prediction algorithms, such systems are currently controversial.

3.2. Analysis of the Legitimacy Basis of AI Intervention in Criminal Justice

The involvement of artificial intelligence in criminal justice is not a simple technological upgrade, it is related to the reconfiguration of power structure, constitutional principles and basic rights. Judicial activities are highly public and authoritative, and the intervention of technical means, especially algorithmic systems, must be carefully demonstrated within the framework of legitimacy. Legitimacy does not only refer to "formal legality", but also "substantial legitimacy", that is, whether AI intervention is clearly authorized by the Constitution and the law, and whether it conforms to the basic spirit of procedural justice and rights protection. This chapter will analyze from three dimensions: constitutionality, legitimacy and legal authorization, and clarify the boundaries of AI judicial development.

3.2.1 Constitutional Analysis

Constitutionalism is the primary prerequisite for AI intervention in criminal justice. Constitutional analysis includes two aspects: one is whether AI intervention challenges constitutional structural principles, such as the principle of judicial independence; the other is whether it infringes on the basic rights of the defendant, including the right to equality, the right to privacy, the right to a fair trial, etc.

The tension between the principle of judicial independence and technological intervention. Article 126 of the Constitution of the People's Republic of my country stipulates that "the people's courts shall exercise judicial power independently in accordance with the law and shall not be interfered with by administrative organs, social groups or individuals." The independent operation of judicial power is the foundation for maintaining judicial authority, fairness and efficiency. However, when AI technology is deeply embedded in the judicial process, algorithm design, model training and operation mechanisms are often dominated by administrative departments or technology companies, which in fact have an impact on the judges' freedom of judgment and subtly change the logic and results of judgment. For example, in the sentencing assistance system, although the sentencing range and case recommendations provided by the system are advertised as "for reference only", in actual

operation, some judges tend to rely on algorithm suggestions to avoid responsibility, thereby substantially restricting the freedom of judgment. If the system results become the "trial standard", it may constitute external interference in the judges' judicial power and conflict with the principle of judicial independence. In addition, once the technology platform is concentrated in a few platforms or controlled by state organs, there is also a risk of "power turning into the host", that is, administrative power uses technology to infiltrate the judicial discretion space, resulting in an imbalance of power. Potential erosion of the basic rights of the defendant. AI systems are involved in the criminal trial process, especially in the areas of evidence assessment, recidivism prediction, and similar case push, which often involve the collection and analysis of the defendant's data, which objectively affects their basic rights:

Privacy rights are limited: AI needs to rely on a large amount of personal data, such as behavioral trajectories, social records, physiological characteristics, etc. for analysis, and whether the collection of these data is authorized and limited directly affects the protection of the defendant's privacy rights. According to the "Personal Information Protection Law" and the "Data Security Law", judicial organs should handle citizen data "minimum necessary", but the implementation of this principle is relatively weak in practice.

Equality rights are violated: Some algorithms may have built-in "bias" or imbalances during the training process. For example, the recidivism risk prediction system may generate "high-risk" labels for specific age groups, ethnic groups, or occupational groups, leading to the problem of "conviction by data". This unequal evaluation based on algorithm output is contrary to the principle of personal equality in criminal law.

The right to a fair trial is diluted: The "Constitution" and the "Criminal Procedure Law" clearly guarantee the defendant's right to an independent and fair trial. If the adjudication process lacks transparency and the system decision cannot be questioned after AI intervention, the defendant's right to defense and right to participate in the procedure will be weakened. For example, the "black box judgment" contained in AI suggestions cannot be questioned and explained, and it is difficult for the defendant to object to the algorithmic bias, which may undermine procedural justice.

3.2.2 Legitimacy Analysis

If constitutionality is an institutional premise, then legitimacy is the normative basis at the practical level. Whether AI intervention in criminal justice is "worthwhile" and "should" depends not only on whether it is effective, but also on whether it meets the essential requirements of judicial justice.

Whether the efficiency advantage of AI is sufficient to constitute the basis of judicial legitimacy. The introduction of AI in the judicial system often takes "improving efficiency", "saving costs" and "unifying standards" as its main selling points. Problems such as tight judicial resources, a surge in cases, and insufficient judicial personnel have made efficiency the core consideration of judicial reform. However, justice that is only efficiency-oriented often finds it difficult to balance "case justice" and "procedural justice". Legitimate justice must not only be "fast", but also "accurate", "fair" and "transparent". If AI technology sacrifices procedural openness, party participation and individual judgment space while pursuing high efficiency, the "efficiency" it achieves may be "pseudo-efficiency". Therefore, technical means must serve the goal of justice, and cannot put the cart before the horse and let technical logic dominate judicial logic.

Whether AI can become an "agent of state power". Whether AI can exercise judicial power on behalf of the state is a fundamental issue. In theory, state power must be clearly authorized by the Constitution and exercised through subjects prescribed by law. Although AI can be understood as an "auxiliary tool", its "depersonalized" technical subject attributes are difficult to be competent for the role of public power responsibility in the context of its increasing decision-making ability. In particular, when the AI system participates in judgment, sentencing or predicting recidivism, its output has a substantial impact on the legal fate of the defendant. In this case, if the deviation or error of the algorithm system cannot be held accountable, the judicial responsibility mechanism will be blank. Therefore, at this stage, we should adhere to the constitutional principle that "AI does not have independent subject qualifications and cannot become the actual bearer of public power" to ensure that any judicial decision is ultimately made by a natural person with legal responsibility.

3.2.3 Legal authorization basis

On the basis of constitutionality and legitimacy, AI intervention in judicial procedures must also have clear legal authorization. According to the Outline for the Implementation of the Construction of a Rule of Law Government and the basic principle of "nothing can be done without legal authorization", all state organs must be authorized to exercise their powers according to law, especially in criminal justice.

Blanks and ambiguities in the Criminal Procedure Law on the application of AI technology. my country's current Criminal Procedure Law has not yet made systematic regulations on the use of technical means such as artificial intelligence in judicial procedures. Only individual provisions involve technical investigations and electronic data acceptance, but there are no

special chapters or sections on functions such as AI-led decision-making, intervention in sentencing, and risk scoring. This has led to a lack of unified standards for the use of technology by courts and procuratorates in practice, and there are problems such as "technical overreach" and "quasi-administrative operation".

Taking "recidivism risk prediction" as an example, there is currently no legal provision that stipulates that the court can decide whether to grant parole or probation based on the system score, nor does it stipulate the review mechanism and objection procedure of the scoring model. This legal ambiguity provides room for judicial arbitrariness.

Whether separate legislation or special authorization should be made. With the increasing application of AI in the judicial field, it has become an inevitable trend to establish a legal system with strong pertinence, clear procedures, and clear rights and responsibilities. Two paths can be considered: a separate legislative path: such as formulating the "Law on the Application of Judicial Artificial Intelligence" or the "Law on Technical Assistance in Criminal Justice", which specifically stipulates the applicable principles, technical standards, review mechanisms, relief procedures, etc. at each stage of AI intervention, forming a complete technical rule of law framework; a special authorization path: by revising the "Criminal Procedure Law", the use of AI is explicitly authorized in specific articles, such as adding "technical assistance clauses" and "algorithm evidence clauses", and limiting the scope of authority of AI-assisted investigation and auxiliary adjudication.

3.3. Analysis of the legal limits of AI’s involvement in criminal justice

The introduction of artificial intelligence has undoubtedly provided efficiency improvement and professional support for criminal justice, but it has also brought many legal and ethical risks. If the “limits” and “irreplaceability” of AI technology are ignored and its power is allowed to expand, it may not only damage the basic rights of the defendant, but also reshape the judicial power structure and shake the foundation of the rule of law. Therefore, it is necessary to clarify its legal limits while promoting the judicial application of AI.

3.3.1 Judicial discretion cannot be replaced by technology

The complexity of free conviction and fact finding. The essence of judicial discretion lies in “free conviction”, that is, judges make judgments through comprehensive analysis of evidence, logical reasoning and emotional considerations. Criminal cases are particularly complex, often involving multi-dimensional factors such as evidence flaws, the psychology of the parties, social background, and the severity of guilt. These factors are highly individualized

and dynamic, and are difficult to incorporate into the “standardized model”. Although AI technology is good at identifying patterns and summarizing similar cases, its algorithm logic relies on established variables and historical data training, and its ability to handle “exceptional cases” and “gray areas” is limited. For example, the judgment of "non-quantitative factors" such as the defendant's remorse attitude, family background, and mental health status is extremely difficult to accurately quantify through algorithms, but it is an important basis for judicial discretion. In addition, criminal trials are not just factual judgments, but also involve legal interpretation and normative application. The law itself is open, vague, and evolving, and it is difficult to cover all legal value judgments simply by relying on technical reasoning. Justice must not only be logical, but also reflect human emotions and ethical intuition.

The complexity of cases exceeds the capabilities of technical models. In recent years, AI has been used for functions such as "similar case push" and "sentencing assistance", which seem to have achieved results in unifying judgment standards and improving judgment efficiency, but in fact there are risks of "oversimplification" and "de-individualization". Especially in some cases with complex plots and obvious conflicts of evidence, the suggestions provided by AI models often ignore the deep structure of the case, and even recommend "similar cases but not similar cases".

For example, a local court tried to introduce "sentencing range recommendations for similar cases" in theft cases, but failed to effectively identify different situations such as "habitual offenders", "recidivists", and "accomplices", resulting in a lack of pertinence in the suggestions. Once this approach is unconditionally adopted by the judge, it may harm individual justice.

Emphasis on the "irreplaceability of judicial personnel". Judicial power is not only the "power of judgment", but also the "power of responsibility". In the case of AI's participation in adjudication, once there is an error or dispute in the adjudication result, how to determine the responsible party? At present, the AI system does not have legal personality and cannot bear legal responsibility, and the legal effect of its output suggestions is not clear. Therefore, the judge should still bear full responsibility for the final adjudication result.

Justice is a highly humanistic activity that requires the moral responsibility and social responsibility of the judge, rather than just the product of "formulaic calculation". AI can be used as a reference tool, but it should not replace human discretion. Adhering to the "people-

centered" judicial value system is the key to preventing technological tools from alienating judicial power.

3.3.2 Algorithmic bias and black box problem

Data bias: How historical discrimination is reproduced in algorithms. The "intelligence" of AI systems essentially comes from learning from historical data. However, these data themselves may contain institutional bias or social discrimination. Once the algorithm is directly applied to the judicial field without cleaning, it may "legalize" past inequality. Taking the "recidivism risk scoring system" as an example, if the training data mainly comes from the high crackdown rate on certain specific groups of people (such as a certain ethnic group or a certain poverty-stricken area), the system may "infer" that these groups are "more likely to recidivate", and then "score and punish" them in the prediction, forming a "label trap" [7].[7] This data bias not only violates the principle of equal rights, but is also likely to form a "self-fulfilling prophecy": the system predicts recidivism → the judicial organs strengthen crackdowns → the data confirms the high recidivism rate → further predicts recidivism, forming a vicious cycle of "technology-enhanced discrimination".

Black box problem: the risk of lack of explainability. AI systems often use complex algorithm structures such as deep learning, forming a "black box mechanism" in the reasoning process, that is, the output results cannot be fully understood or explained by humans [8]. In criminal justice, if the basis for the judgment comes from an inexplicable system output, the defendant and his defense counsel will lose the ability to question the judicial process, thus shaking the foundation of procedural justice. For example, the "COMPAS Risk Assessment System" case in the United States has caused widespread controversy [9]. The system evaluates the defendant as "high risk", which directly affects the judge's judgment on sentencing and parole, but the defendant cannot know the basis and calculation logic of the score, nor can he make a targeted defense. The case was exposed by the media, triggering a dispute over "algorithm transparency" and "verifiability". In the end, although the court recognized the rationality of the use of AI, it also questioned the legality of "black box judgment".

3.3.3 Procedural justice and protection of the right to defense

Can algorithmic evidence be cross-examined? The core spirit of criminal procedure is equality between prosecution and defense and procedural fairness. However, AI intervention often reflects its role in the form of "evidence", such as "risk prediction score", "public opinion model results", "behavior trajectory analysis", etc. Whether this algorithmic evidence can be cross-examined is directly related to the realization of procedural justice. At present, in

practice, defendants often find it difficult to obtain information about the model principles, algorithm processes, data sources, etc. of AI systems, which makes it impossible for them to raise substantive questions about the system output, forming a situation where "technology cannot be challenged". This phenomenon essentially deprives the defendant of the right to cross-examine and the right to defense [9].

The gap in the technical capabilities of the defense. Even if the defense is given the opportunity to cross-examine, whether it has the ability to fight against complex algorithms is also a realistic problem. Traditional criminal defense lawyers are mainly engaged in legal analysis and fact investigation, with weak technical backgrounds. When facing algorithmic systems, they often "have the right to cross-examine but not the ability to cross-examine". The inequality of technology can also easily aggravate the imbalance of judicial resources, making AI a "technical helper" for the prosecution. Therefore, a technical expert assistance system should be established to provide technical interpretation support for the defense, and at the same time, law schools and bar associations should be encouraged to set up "technical legal training courses" to improve the "digital literacy" of defenders [10].

3.3.4 Data abuse and privacy protection

The boundaries of criminal database use are blurred. In the process of AI technology-assisted investigation, sentencing, and prediction, data resources have become its core support. Public security, procuratorates, and courts in various places have established a large number of criminal databases, including face recognition databases, call records, and social media analysis databases. However, the boundaries of the use of these databases are still unclear, such as frequent use outside of the purpose: databases originally used for criminal cases are used for public security, administrative law enforcement, and even commercial behavior, violating the principles of data minimization and purpose limitation; data retention period is too long; long-term retention of data of those who have not been convicted or have reformed may constitute "invisible punishment"; lack of exit mechanism: individuals find it difficult to apply for deletion and modification of false data [11].

The applicability and shortcomings of the "Personal Information Protection Law". The "Personal Information Protection Law" establishes the principles of legality, legitimacy, necessity, and good faith, requiring that the processing of personal information should have a clear purpose, obtain consent, and protect rights. However, in criminal justice, data use is often excluded from the "consent" principle, especially during the investigation stage, when judicial authorities can collect personal data without authorization. This "purpose first"

judicial exception is reasonable to a certain extent, but if there is a lack of supervision and restrictions, it is very likely to lead to data abuse. Moreover, the law's supervision mechanism for "algorithm judgment" is not clear, and the requirements for "assessment of the impact of algorithms on human rights" are not specific, resulting in limited binding force at the implementation level [12].

4.Discussion For Suggestions on the establishment of a system to regulate AI’s involvement in criminal justice

With the rapid development of artificial intelligence technology, the application of AI in criminal justice has become an irreversible trend. How to strike a balance between promoting judicial efficiency and ensuring judicial fairness, and avoiding technological abuse and alienation of justice, lies in the response at the institutional level [14]. We should start from the four dimensions of legislative improvement, role positioning, technical supervision and rights relief, and establish a systematic, scientific and forward-looking AI judicial governance framework.

4.1. Improve legal authorization and governance system

4.1.1. Revise the Criminal Procedure Law

At present, my country’s Criminal Procedure Law has not yet made clear regulations on AI’s involvement in criminal procedures. Related applications mostly rely on departmental regulations, local pilot projects and administrative promotion, and lack a unified legal framework. In order to respond to the institutional gap brought about by technological development, it is recommended to set up a special "smart justice" chapter in the Criminal Procedure Law to clearly stipulate the scope of application, procedural guarantees, and responsibility mechanisms of AI applications:

Scope of application limitation: It is explicitly stipulated that AI can be used for "auxiliary links" such as evidence screening, similar case recommendation, and risk prediction, and cannot independently make judgments or replace the judge’s discretion [15].

Rights protection clause: requires that the use of technology should respect the parties' right to know, right to cross-examination and right to relief, and ensure procedural fairness and equal rights. Technical specification review: clarifies that AI systems must be reviewed and approved by the national judicial technology certification agency before they can enter the judicial process, ensuring their scientificity and legality. This legislative measure helps to

establish the legal basis of AI judicial activities at the source and avoid the institutional dilemma brought about by "technology first, law lags behind" [16].

4.1.2. Establish a special regulatory agency for the judicial use of AI

Technology application involves cross-domain and multi-departmental collaboration, and the current judicial system is difficult to independently undertake the systematic task of AI governance. It is recommended to establish an independent "Artificial Intelligence Judicial Application Supervision Committee", led by the Supreme People's Court and jointly composed of the procuratorate, the Bar Association, and the science and technology department. Its responsibilities include: formulating AI judicial technology application standards; reviewing the qualifications of AI systems to enter the judicial process; evaluating the risk level of AI use in local judicial organs; accepting public complaints and reviewing technology abuse. By institutionalizing the establishment of a "third party for technical supervision", it can effectively avoid the role conflict of a single agency "both as a user and an evaluator", and enhance the credibility and authority of AI governance.

4.2. Clarify the boundaries of AI roles

4.2.1. Judges' inalienable power of adjudication should be explicitly protected by law

The essence of technical tools is "means", while judicial adjudication is a "value judgment" based on multiple considerations of law, humanity, ethics and society. Therefore, the bottom line that "judicial adjudication power belongs exclusively to judges and cannot be replaced by technical systems" should be clarified at the legal level [17]. All AI system's adjudication suggestions are only "reference materials" and do not have independent legal effect; judges should independently review, interpret and select the suggestions provided by the system and explain the reasons for their judgment in the adjudication documents; technical suggestions should not be used as the basis for "adjudication automation" to prevent the formation of the phenomenon of "procedural substitution responsibility". This specification aims to prevent the trend of "people giving power to machines" and to maintain the core values of judges' subjectivity, responsibility and judgment.

4.2.2. Technology should only exist as "auxiliary evidence"

In principle, information generated by technology, such as behavioral trajectory prediction, voice analysis results, social risk maps, etc., should be included in the category of "auxiliary evidence". Its probative value is lower than that of original evidence and personal and physical evidence. The applicable conditions include: it can only be used for corroborative

and clue functions, and cannot be used alone as the basis for conviction or sentencing; it must be subject to the defense cross-examination procedure; when there is reasonable doubt or explanation is impossible, the "favorable to the defendant" judgment principle should be applied. By clarifying the "evidence level of technical evidence", the problem of AI evidence "taking the lead" in criminal proceedings can be effectively avoided [18].

4.3. Establish an algorithm disclosure and explainability mechanism

4.3.1. The government and enterprises need to disclose the algorithm logic and data source

At present, AI judicial systems are mostly developed by enterprises or local administrative agencies. Their algorithm models, training data, weight distribution, etc. are mostly "commercial secrets". The public and judicial personnel often find it difficult to understand their reasoning process, forming a "black box judgment". To prevent abuse of power, judicial AI systems must be required to report their algorithm structure and logic description to the court and the procuratorate; key systems (such as sentencing assistance and risk prediction) must be forced to disclose their model files, variable settings and data sources; and third-party experts must regularly conduct algorithm risk assessment reports and accept social supervision. Only by enforcing "algorithm transparency" by law can a reviewable basis for the operation of judicial power be established.

4.3.2. Defendants should enjoy the "right to know about algorithms" and "right to question"

As the most direct rights bearer in the procedure, defendants should enjoy rights protection. Including: the right to know, the defendant and his or her defense counsel have the right to know the analysis results, operating principles and data basis of the AI system for this case; the right to question, if the defendant has objections to the system's suggestions or evidence, he or she should have the right to apply for re-evaluation or hire experts to conduct adversarial analysis; the right to remedy, the court should allow the application for evidence exclusion due to algorithm bias or reasoning defects to protect the defendant's legitimate rights. To protect such rights, judicial organs should provide defense counsel with algorithm interpretation toolkits and basic training, and promote the establishment of a cross-disciplinary expert database [19].[19] At the same time, the principle of "explainability" should be strengthened during the AI system design phase, and algorithm structures with logical backtracking functions should be promoted.

4.4. Improve the data protection system and citizen relief mechanism

4.4.1. Strengthen the classification management mechanism for criminal data

There are many types of data in the criminal justice field, involving sensitive information such as personal identity, behavior records, communication records, social interactions, and network behavior. It is recommended that on the basis of the existing "Data Security Law" and "Personal Information Protection Law", criminal data be refined into categories such as "identifiable data", "linkable data", and "non-sensitive behavior data"; different access rights, retention periods, and usage conditions should be set for different categories of data; and the "principle of minimizing the use of judicial data" should be established, emphasizing that how much data is used, how long it is retained, and who can use it must be legally clear [20]. In addition, the "criminal data desensitization use" technology, such as data falsification and differential privacy, should be promoted to protect individual privacy.

4.4.2. Establish an "algorithm harm relief channel"

If the defendant's rights are damaged due to errors, biases, or abuses of the AI system, the current legal relief path is relatively vague. It is recommended to establish a "three-way relief mechanism for algorithm harm". For example, administrative reconsideration allows parties to file a technical review request with the regulatory agency when AI applications produce data errors or program flaws; judicial review allows courts to accept objection lawsuits against the application of AI systems and requires relevant system developers to bear the burden of proof; public interest litigation supports the prosecution or social organizations to file public interest lawsuits to promote system rectification when the system is generally biased, infringes on privacy, or affects the rights of groups. At the same time, an AI judicial use liability insurance system should be established to guarantee economic compensation for technical damage liability.

References

[1] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671–732. https://doi.org/10.2139/ssrn.2477899

[2] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency*, 149–159. https://doi.org/10.1145/3287560.3287583

[3] Chen, R. (2020). Institutional dilemmas and rule-of-law responses to the intervention of artificial intelligence in criminal proceedings. *China Legal Science, (6)*, 5–26.

[4] Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review, 89*(1), 1–33. https://doi.org/10.2139/ssrn.2376209

[5] Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press. https://doi.org/10.2307/j.ctt1zxxj6t

[6] Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

[7] Hao, K. (2019). AI is sending people to jail—and getting it wrong. *MIT Technology Review*. https://doi.org/10.5281/zenodo.4456090

[8] Kroll, J. A., et al. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705. https://doi.org/10.2139/ssrn.2765261

[9] Koops, B. J. (2010). Law, technology, and shifting power relations. *International Review of Law, Computers & Technology, 24*(1), 7–15. https://doi.org/10.1080/13600861003673273

[10] Lepri, B., et al. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology, 31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

[11] McGregor, L., Murray, D., & Ng, V. (2019). International human rights law as a framework for algorithmic accountability. *International and Comparative Law Quarterly, 68*(2), 309–343. https://doi.org/10.1017/S0020589319000046

[12] Mittelstadt, B. D., et al. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

[13] Pasquale, F. (2015). *The Black Box Society: The secret algorithms that control money and information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674915660

[14] Wang, G. (2021). Risk types and institutional responses of artificial intelligence participating in judicial adjudication. *Chinese Journal of Law, (5)*, 125–142.

[15] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law, 7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

[16] Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance, 12*(4), 505–523. https://doi.org/10.1111/rego.12160

[17] Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of Criminology, 18*(5), 623–642. https://doi.org/10.1177/1477370820941397

[18] Zarsky, T. Z. (2016). The trouble with algorithmic decisions. *Science, Technology, & Human Values, 41*(1), 118–132. https://doi.org/10.1177/0162243915605575

[19] Zeng, J., Lu, Y., & Huang, M. (2022). Algorithmic justice in Chinese courts: Promise and perils of intelligent trial systems. *Information, Communication & Society, 25*(6), 829–845. https://doi.org/10.1080/1369118X.2020.1864006

[20] Zuo, W. (2023). Ethical governance of AI in criminal justice: Lessons from China's smart courts. *AI & Society*. https://doi.org/10.1007/s00146-023-01599-9