DOI: 10.23977/law.2024.030324
ISSN 2616-2296 Vol. 3 Num. 3

Application of Synthetic Data in Artificial Intelligence Trials from the Perspective of Judicial Justice

Rui Shen[*]

BTBU School of Law, Beijing Technology and Business University, Beijing, China
15621507637@163.com
[*]*Corresponding author*

Keywords: Synthetic Data, Judicial Artificial Intelligence, Judicial Justice

Abstract: China is actively developing its intelligent court system, which incorporates artificial intelligence technology into judicial trial proceedings. The aim is to alleviate the burden of a high number of cases and personnel shortages while also enhancing the quality and efficiency of the trial process. However, this initiative also presents several drawbacks, which pose significant challenges to attaining judicial justice. The development of artificial intelligence relies on training data, which can lead to defects in AI systems due to factors such as inadequate quantity, lack of structure, algorithmic discrimination, algorithmic black box, and privacy leakage. Introducing synthetic data into artificial intelligence trials is critical because of its low cost, diversity, and good security, which can compensate for judicial data's limitations in various ways. We can successfully avoid the hazards associated with existing artificial intelligence in trial activities by substituting the original judicial data with synthetic data, thereby enhancing digital justice and judicial fairness.

1. Introduction

Traditional courts lack intelligent trial assistance and struggle to meet the high demand for judicial proceedings. This inadequate judicial supply poses a significant threat to judicial impartiality. In order to solve the reform issues, the Supreme People's Court suggests that we push for advancing science and technology, put the network strengthening strategy into practice, completely construct intelligent courts, and make full use of contemporary scientific and technological tools like big data, cloud computing, artificial intelligence, etc. Courts at all levels are actively encouraging the development of intelligent court system to address the issue of an overwhelming number of cases and limited personnel, as well as to improve the standard of justice. Trial proceedings have also incorporated the implementation of intelligent systems. Artificial intelligence in the legal field comes in many forms, but the most common ones are electronic case information, the intelligence of the system that handles cases, predicting and recommending important decisions, standardizing evidence criteria, and other specific uses[1]. With the assistance of various policy documents, local courts are actively promoting the development of intelligent court system. Beijing, Shanghai, Jiangsu, and Anhui have already implemented judicial artificial intelligence systems. As an illustration, the courts in Beijing have introduced the "Judge Rui" system, which is an intelligent research and judgment system. Similarly, the courts in Shanghai have implemented the "206" system, which is an intelligent auxiliary case-handling system specifically designed for criminal cases. Additionally, the courts in Suzhou have established the "Suzhou Model of Intelligent Trial".

In the age of information technology and digitalization, the significant increase in data serves as a solid foundation for the rapid advancement of artificial intelligence. The trend toward the use of intelligent systems in the judicial field is unstoppable. Insufficient judicial data, both in terms of number and quality, impedes the AI trial system's progress and poses a concealed risk for the emergence of associated dangers. The primary purpose of a judicial trial is to uphold justice. To achieve this, it is crucial to address the current challenges faced by artificial intelligence trial systems and enhance their ability to assist judicial personnel in conducting fair trials. In doing so, we can make justice more tangible and accessible.

2. The Necessity and Realistic Dilemma of AI Applied to Judicial Trials

2.1. The Necessity of Applying AI in Judicial Trials

2.1.1. Intelligent Trial to Improve the Judge's Trial Efficiency

In recent years, there has been a significant increase in the number of cases handled by the people's courts due to increased public awareness of the rule of law. In 2022, the local people's courts handled a total of 33.704 million cases at all levels, including specialized courts. On average, each judge at the grassroots level handled more than 300 cases per year. Aside from handling essential trial matters, judges frequently encounter a significant volume of monotonous and exceedingly time-consuming auxiliary tasks, such as organizing and inputting case files, submitting and delivering information, and proofreading documents. The increasing workload of judges necessitates the need for more intelligent and efficient tools to aid them in dealing with non-judgmental concerns. This will help prevent hasty conclusions and ensure a thorough understanding of each case. According to Posner's theory in his book *The Economics of Justice*, the role of law is to enable the realization of efficiency, and improving efficiency generates the circumstances necessary for the fulfillment of justice. Postponed justice is not justice at all. Within the context of the "many cases, few people" predicament in court, judicial trial proceedings must efficiently utilize limited judicial resources to achieve optimal justice. Artificial intelligence functions based on predefined algorithms and excels at performing numerous monotonous tasks with greater precision and efficiency than humans. For instance, it can automatically arrange trial outlines and summarize key arguments, eliminating the need for judges to repeatedly review case files and input data. By reducing the workload of administrative tasks, the judges may allocate more time and effort to the actual trial proceedings and render decisions with careful consideration. We can use artificial intelligence to allocate judicial resources rationally and accurately, leading to improved trial quality and efficiency. This allows for prompt responses to the public's judicial needs.

2.1.2. Intelligent Trial to Break Information Barriers

The traditional trial model stores and exchanges judicial information and resources internally across different regions, levels, and departments. Despite the implementation of electronic case handling by each, the independent development of their systems' technical architecture and data standards leads to a lack of uniformity. As a result, there are significant obstacles to the efficient flow and utilization of judicial data. The presence of data fragmentation and hurdles within the court system can easily lead to judicial inconsistency as a result of the absence of standardized data standards. To effectively utilize the data, an artificial intelligence system can integrate scattered data resources from multiple directions and analyze them using various data mining techniques. This promotes the sharing of judicial data resources and facilitates collaboration between departments, thereby addressing the issue of "data islands" to some extent. Using the case recommendation function as an example, the algorithm can use a lot of judicial data to examine the relationship between a case's facts and result by gathering judicial data into a database. Judges input the elements of a case, and artificial intelligence uses the correlations from similar cases to recommend a sentence. The initial judicial data comes from various departments, courts. Ultimately, this system benefits the entire political and legal system by facilitating the sharing of information among departments and courts. The use of artificial intelligence facilitates the efficient use of data resources, dismantles data obstacles among different departments, regions, and levels of the political and legal system, and enhances the standardization and impartiality of judicial trials.

2.1.3. Intelligent Trial to Optimize the Regulatory Process

In the past, the responsibility for overseeing and rectifying errors in the trial process rested on the dean and the president, who had to proactively intervene. However, the boundaries between supervision and management were not clearly defined, resulting in the dean and president frequently being passive and unwilling to supervise. Additionally, due to the court's overwhelming caseload and limited personnel, the dean and the president faced energy constraints that made it difficult to effectively supervise every stage of the trial process. As a result, the majority of supervision only occurred after the case review through spot checks[2]. The level of digitization in the trial process is insufficient, resulting in only the case officer having a comprehensive understanding of the specific case information. It is challenging for other supervisory bodies to obtain accurate and complete case information, which consequently hinders the timely reminder and correction of non-standard procedures. The trial process involves multiple nodes, and the lack of comprehensive supervision can result in the inability to identify the specific stage where an error occurred, leading to confusion when attempting to remedy the faults afterward. The Supreme People's Court has recommended that all stages of the litigation process, from the filing of a case to its conclusion, be incorporated into the information-based case handling platform. This will ensure that a comprehensive record is maintained throughout the process. The trial process seamlessly incorporates artificial intelligence, offering comprehensive coverage from pre-litigation to post-judgment. It automatically detects and alerts for any irregularities or flaws and, through intelligent analysis, identifies and reports defects to the dean and the president, prompting them to promptly oversee the situation. By utilizing the intelligent system's reminder, trial staff, and supervisors can promptly identify and rectify errors, effectively mitigating the drawbacks associated with delayed monitoring. Artificial intelligence not only governs the conduct of court professionals but also facilitates the uniformity of the trial process and enables judicial activities to run transparently.

2.2. The Realistic Dilemma of AI Applied to Judicial Trials

The primary obstacle confronting the existing artificial intelligence trial system is its lack of advanced intelligence, particularly evident in its inability to accurately discern facts, thereby increasing the likelihood of wrongful convictions. Data is considered the fundamental basis for the advancement of artificial intelligence, and the progress of artificial intelligence encounters significant challenges, some of which arise from diverse imperfections in the training data. The training data for judicial AI is made up of authentic judicial data, including trial execution data, judicial administration data, legal papers, and other types of data. Overall, various forms of judicial data commonly encounter issues such as inadequate quantity and low degree of structuring, which directly hinder the efficient and high-quality development of judicial AI.

2.2.1. Inadequate Magnitude of Judicial Data

The judicial data used in the development of the current artificial intelligence trial system primarily originates from the online document data of the China Judgment and Decision Network. By the end of 2023, the adjudication document network will have accumulated more than 100 million documents. However, although the current number of available documents is significant, there is still a substantial gap in achieving the goal of "massive and high-speed" judicial big data. Furthermore, this gap hinders the provision of favorable support for the advancement of artificial intelligence. First, the public's access to judicial documents began in 2013. This implies that we cannot effectively use any non-electronic data from before 2013. Secondly, due to existing technical and institutional limitations, the internet does not promptly make judicial documents accessible, leading to slow updates. This significantly hampers the amount of judicial data available. To ensure effective administration of justice, it is necessary to enhance the intelligence of the intelligent system. However, the current volume of judicial data is insufficient to meet the requirements for system upgrading. Upgrading the entire court data is challenging to accomplish quickly, but the resolution to the technical obstacles is forthcoming, so creating a paradox.

2.2.2. Monopolization of Judicial Data

Judicial data about sensitive social issues, juvenile crimes, and other violent and graphic content exists in significant quantities. However, judicial authorities frequently opt to withhold or only partially disclose this data due to concerns regarding personal privacy, national security, and societal repercussions. Typically, the public closely monitors information about these cases. In the realm of digital justice, the state holds a dominant position in terms of data control, and state authorities gather information to fulfill their role in social governance. Over time, they establish data warehouses, conduct online analysis, and engage in data mining[6]. Revealing only some judicial data will unavoidably result in an imbalance of trial knowledge between the government and the people. Information asymmetry undermines the principle of equal confrontation in the realm of criminal proceedings. The defendant and their defense lawyers cannot access and analyze data, making it challenging to obtain resources and information technology on par with those available to national public authorities[7]. The defense faces challenges in creating an effective defense, and the disparity in resources between the prosecution and the defense threatens the structural integrity of a fair trial and raises the likelihood of a wrongful conviction.

2.2.3. Low Degree of Structuring of Judicial Data

Constructing a knowledge graph is a necessary process for AI training. Judicial data is generally less structured, which increases the difficulty of graph construction. A multitude of factors in the case necessitate manual categorization and pre-summarization, resulting in increased labor costs. In addition, the scope of the cause of action of cases is limited. There are over 400 categories of causes in criminal cases, and the causes of action of civil cases can be further divided into thousands[9]. However, the "Shanghai Intelligent Aid Case Handling System for Criminal Cases" only includes 18 specific crimes, and the intelligent trial model used by the Guizhou court only covers common causes such as intentional injury, robbery, and theft in criminal cases[9]. Consequently, the Low degree of structuring of judicial data is the primary reason for excessive dependence on manual processing of judicial data, which is also the major challenge encountered by AI trials currently.

3. Enhancements of Synthetic Data for AI Trial Systems

3.1. Overview of Synthetic Data Concepts and Characteristics

Synthetic data refers to artificially generated data that closely resembles real-world data in terms of statistical properties[3]. These data are created using algorithms to replicate real-life circumstances. Synthetic data is produced using specific algorithms and models to replicate the characteristics and distribution of actual data. Although synthetic data does not consist of real data, it closely resembles real data and more numerous than the real data. This unique quality enables synthetic data to substitute real data in the advancement of artificial intelligence, thereby expediting the development process and diminishing associated expenses.

Synthetic data is characterized by its low cost. Before obtaining high-quality data, it is important to preprocess the data, which involves addressing missing data and anomalous data, among other factors. To ensure the model training meets expectations, manual labeling of each data item is required, significantly raising data acquisition costs and limiting dataset scalability[3]. Generating synthetic data is less laborious compared to working with actual data. The synthetic data has been labeled during the data generation process, allowing for the rapid generation of hundreds of text samples that meet the specified criteria. The labels for all the text samples are also prepared during the generation process, resulting in significant cost savings for subsequent data collection, labeling, and management tasks. Certain producers of synthetic data emphasize that the expense of manually labeling an image is $6, whereas manual synthesis only costs 6 cents[4].

Synthetic data exhibits greater diversity. During the trial process, it is necessary to address diverse scenarios; nevertheless, there are variations in judicial data for each scenario. As of November 2023, 61% of the total number of documents in the China Judicial Documents Network are civil records, while just 7% are criminal documents. Synthetic data refers to data that is artificially made by humans. It was created to produce a greater quantity and wider range of data, using existing legal data as a basis. This helps address the issue of limited and unbalanced data resources while also offering a larger pool of high-quality data for training artificial intelligence. In actuality, there exists a certain circumstance where it is inherently challenging to collect authentic data. For instance, while simulating the training of artificial intelligence for driving, it is not feasible to obtain real data, including all possible road conditions. This limitation gives rise to what is known as the "long tail effect." Synthetic data can address the lack of edge cases and achieve wider data diversity. The use of synthetic data in judicial trials can address the shortcomings of data imbalance, make up for and supplement missing and relatively limited judicial data, seek to cover a wide range of judicial scenarios, prevent the biased treatment of specific groups, and maintain the fairness and impartiality of the judicial system.

Synthetic data offers enhanced security and confidentiality. Synthetic data generates data that is similar to real data but does not contain real personal information by utilizing different techniques, which also dictates that the process of data analysis, model development, and algorithm testing can be performed without exposing real data. The purpose of generating synthetic data is to enable the analysis of data, the construction of models, and the testing of algorithms without the need to expose real data. The United Nations Big Data Working Group recently issued *The UN Guide on Privacy Enhancing Technologies for Official Statistics*, which includes 18 prominent examples of privacy-enhancing technologies from throughout the world. Two of these cases employ methods involving "synthetic data." While synthetic data closely resembles real data, it lacks identifiable and traceable information, allowing for broader sharing and utilization compared to genuine data. When compared to other types of data, judicial data has stricter standards for maintaining confidentiality. To ensure the security of judicial data, it is essential to not only make modifications at the institutional level but also to explore technical factors. Utilizing synthetic data that excludes personally identifiable information for system development not only safeguards public privacy but also facilitates the expansion of judicial disclosure. Public scrutiny of judicial behavior and understanding of judicial decision-making is the only way to ensure that power operates in the sunlight and that judicial justice is achieved. Synthetic data can provide more, safer and higher-quality data at lower costs, providing opportunities to fully utilize the value of data elements.

3.2. The Necessity of Incorporating Synthetic Data into AI Trial Systems

3.2.1. Expanding the Scale of Judicial Training Data

Throughout the evolution of ChatGPT, the necessary quantity of training data has experienced a significant surge, escalating from 1.5 billion parameters at the introduction of GPT-2 to a staggering 175 billion parameters employed in the launch of GPT-3. It is evident that AI heavily depends on extensive training data to enhance its intelligence[5]. Without adequate data scale, even increasing computational capacity cannot significantly enhance the intelligence level of AI. Synthetic data has the ability to simulate genuine data, and it can generate a large number of text samples that fulfill specific criteria in just a matter of minutes. Naturally, synthetic data is not generated from nothingness; it still originates from actual data. Specifically, the generation process of synthetic data can be divided into the following steps: first, the extraction of real judicial data samples and data models attached to the label, such as the cause of the case, the region, and so on; the specific types of labels in accordance with the training needs; and second, the data will be placed in a variety of scenarios randomly combined according to the needs of the situation to quickly generate a number of groups of data to meet the requirements. Therefore, in the training of the same AI model, real judicial data can only express a certain type of case referred to by the data, and if you want to broaden the types of cases, you need to screen and accumulate more real data, while synthetic data can be adjusted only by the cause of the case, the region, and other parameters within a minute to produce hundreds of different data sets to improve the efficiency of the data set generation. Even if a given sort of judicial data is extremely rare and difficult to get, the AI's understanding of such cases will not be influenced by the lack of data size. Synthetic data can effectively overcome the problems brought about by the lagging and inefficient uploading of judicial data, satisfy the needs of the current rapid growth of judicial artificial intelligence, and facilitate trial operations.

3.2.2. Promote the Disclosure of Judicial Data

Synthetic data technology can encourage the decentralization of data, make data resources accessible and useful to a larger range of individuals, help solve the problem of data silos, and promote the portability and sharing of data. For example, in the case of certain court records, including brutal and violent material, the judicial authorities have chosen to view them internally exclusively and not divulge them to the public, taking into account that their wide dissemination might have a negative social impact. This method takes into account public feelings, public order, and morals, but at the same time adds to the difficulty of accessing the data. The undisclosed portion of the judicial data will also generate training gaps when the AI is being built, leading to miscarriages of justice when the AI is involved in the trial of such cases in the future. Synthetic data are statistically comparable to, but not part of, real data, which allows the judiciary to exclude constraints and make available to the public the kind of judicial data that must be monopolized by the state.

3.2.3. Provide Highly Structured Training Data

Artificial intelligence automatically mines and anticipates vast volumes of data through machine learning and finally obtains intelligent algorithms or reference guides. The data for AI machine learning must be structured data, and only by feeding AI structured data samples and refining the commonalities and patterns can we make judgments about similar situations that may arise in the future. Taking the class case search function as an example, knowledge graph construction is to establish the relationship between the basic facts, the focus of the dispute, the application of the law, and the ultimate conclusion through technical language. Judicial AI also needs to refine the laws and categorize the data according to diverse case conditions. Through deep learning of these facts, the computer builds human-like information extraction and logical analysis capabilities, which in turn support the referee[8].

Judicial data is mainly in unstructured and semi-structured form, so the construction of the knowledge graph cannot be done "bottom-up" because artificial intelligence can only be based on pre-set labels to extract the information in the entity and cannot autonomously and accurately complete the information extraction work[8]. Synthetic data is generated based on the statistical properties of real data, i.e., in the production process, synthetic data has already been self-labeled, and it may meet the demand for structured data for AI training without artificial aid in labeling. Therefore, the wide number of synthetic data and the high degree of structure create conditions for the construction of the knowledge graph in a "bottom-up" way, so that we can get rid of the construction method of setting up categories and then matching according to the categories, which not only reduces the dependence on manual labor but also expands the coverage of the case.

3.3. Synthetic Data Solves Attribution Challenges

The development and application of an artificial intelligence trial system have a multifaceted impact on the current judicial accountability system. Most importantly, it blurs the lines between trial rights and responsibilities[7]. Artificial intelligence helps by taking part in the exercise of trial rights, which can change trial rights. AI system development can not be separated from technology outsourcing, in which case the trial power is partially divided, and the trial responsibility is also divided and transformed. Simultaneously, the diversification of the major body of trial responsibilities inherently introduces the issue of responsibility attribution. Judicial data as a data source is gathered by state agencies, but the data must be tagged and processed before being fed to AI, a task that is frequently outsourced to third-party private corporations. The outsourcing of technology development will give rise to a new subject of accountability, influencing the prior straightforward attribution model of "the adjudicator decides, the adjudicator is responsible." The subjects participating in the trial are unknown, and it is impossible to apportion guilt after a miscarriage of justice, which becomes the biggest risk following the implementation of the AI trial system. To answer the problem of how to define the responsibility of judicial artificial intelligence after participating in the trial, it is required to increase the interpretability of judicial artificial intelligence decision-making and its usage of data. In principle, judges have trial responsibility for the cases they undertake, and the answer to the question of who is liable for the miscarriage of justice in cases caused by AI algorithmic errors and algorithmic prejudice is equivocal. Some experts believe that the technicians should bear the responsibility for miscarriages of justice due to technical issues, while the judges should fulfill their duty of care. For instance, when scientific and technological personnel developing the artificial intelligence trial system failed to accurately transform the judicial data code due to technical defects and miscarriages of justice, they should hold the technicians accountable. If the trial staff could have prevented the error by adhering to the duty of care, but chose not to do so, they should also bear responsibility. There seems to be a clear separation between technical responsibility and the judge's responsibility. While assigning blame in this manner makes sense, the process of outsourcing technology, such as modifying code, constructing knowledge graph, or creating algorithms, involves an unknown number of components. The trial rights are all divided into an unknown number of parts, and the workload of each part and its impact on the outcome cannot be calculated. Especially when it comes to specialized issues in the field of artificial intelligence, it is not possible to completely clarify the responsibilities of each part.

Artificial intelligence, regardless of its level of development, cannot completely replace the judge in decision-making. It can only serve as a reference and assistance for judicial personnel during trials, and it cannot arbitrarily divide the right to trial. To solve the problem of attribution of responsibility brought about by artificial intelligence, it is necessary to reduce the excessive dependence on human beings in the training process of artificial intelligence and to prevent the technicians from consciously or unconsciously implanting their own value bias into artificial intelligence. Because synthetic data itself has the characteristic of "labeling", if it is used in the training stage of artificial intelligence instead of real judicial data, the "data source - algorithm generation" stage can eliminate the need for manual labeling of a large amount of real judicial data. In the "data source - algorithm generation" stage, the manual labeling of a large amount of real judicial data can be eliminated. This means that the AI training data can be used directly without the intervention of third-party technicians. When judicial personnel and technicians use real judicial data to train AI, they are likely to shirk responsibilities to each other. The technicians will think that there are defects in the judicial data and transfer the responsibility to the data provider, while the data provider will transfer the responsibility to the technicians on the basis of improper data labeling. Synthetic data not only makes up for judicial data deficiencies and decreases the danger of miscarriage of justice, but also prevents judicial people from passing the burden of refereeing errors to artificial intelligence. Artificial intelligence generates the synthetic data itself. In this scenario, if the judicial AI generates erroneous results due to bias, leading to a miscarriage of justice in a case, it may be necessary for adjudicators of the algorithmic decision-making to assume the responsibility of interpretation. In contrast, the use of real judicial data to train AI is ambiguous in terms of who is responsible, the technicians or the judiciary. Though the intervention of artificial intelligence technology creates a lot of uncertainty, the constitution sets judicial power as a public power that must adhere to clear responsibility constraints. This not only ensures the independence of judicial power but also fosters public respect and trust in the judiciary.

4. Impact of Synthetic Data on the Fairness of AI trials

The everlasting goal of judicial activities, i.e., judicial justice. The application of artificial intelligence must also be grounded in justice. In order to assess whether a court judgement upholds the principle of impartiality, it is necessary to consider the two aspects of impartiality: the fairness of the judicial outcomes and the fairness of the judicial procedures[10]. To actualize judicial justice, it is required to adhere to the fact-based and correct application of the law. Concerning procedural justice, Professor Wang Liming describes in his book six aspects: independence and neutrality of the adjudicator, procedural logic, procedural openness, procedural equality, procedural democracy, procedural convenience, and timeliness. First, AI forms its adjudication conclusions entirely on the basis of the factual characteristics of the case: it analyses the laws based on prior adjudication and forecasts the outcome of comparable instances, which is vital for achieving the same verdict in the same case and judicial unity. Second, the incorporation of AI into trial activities increases the efficiency of judicial work, and the online disclosure of documents meets the public's right to know, helping to realize judicial procedural justice. However, the gradual emergence of data defects, algorithmic discrimination, privacy leakage, and other problems hinder the effective utilization of the AI trial system's advantages in both substantive and procedural areas of justice, potentially raising questions about the impartiality of justice among the public. The process of intelligently building the legal system should not overlook the positive effects of artificial intelligence on the implementation of judicial justice, but it also requires overcoming the resulting challenges.

4.1. Bridging the Gap in Judicial Training Data

What quality of training data will train what quality of AI. In terms of quantity, quality and structure, the current judicial data cannot support the AI trial system for optimization and upgrading, but instead affects the accuracy of the AI system's functions. Due to the imbalance and incompleteness of the data, it can easily lead to algorithmic discrimination and create an algorithmic "black box." According to this tendency's evolution, the deployment of an artificial intelligence trial system will even pose a serious threat to judicial justice. The development of an artificial intelligence trial system must undergo a process that involves processing non-structured or semi-structured raw judicial data, converting it into code, legal language, or natural language, and finally converting it into computer language. In short, it must extract and mark the key information of the statement so that the artificial intelligence can understand the input language. We can only develop a model that accurately links the case facts with the decision outcome by leveraging enormous volumes of labelled data for training. Williams states in *Language and Law* that "the language that constitutes a legal text is more or less always unclear.[11]" The legal language's ambiguity, the difficulty of its transformation, and the error-prone manual integration and refinement of the input data make the already limited judicial data act as a distraction, thereby impairing the accuracy of the AI judgment. Moreover, AI researchers, who largely spearhead the development of legal AI, face the challenge of high technicality and weak legal literacy. Third-party technology businesses typically handle language conversion, manual labeling, and other data product development activities, but they cannot guarantee the legal literacy of the data processor or the accuracy of data analysis and processing. Not designing judicial AI technologies with a legal mindset is likely to pose practical challenges[12]. Particularly, the accuracy requirements for the class case search, class case recommendation, referee deviation warning, and other functions of artificial intelligence are high and have a direct impact on the judge's consideration of the case decision. However, the current artificial intelligence system is primarily plagued by issues such as inaccurate paperwork identification, inaccuracies in search function for class cases, overly broad determinate predictions, and other problems. If the advice provided by AI becomes an interfering factor in the judge's trial process, ensuring justice through AI-assisted same-case same-judgment adjudication will become a pipe dream. At the same time, AI can't cope with unforeseen scenarios as it summarises the facts and results of prior cases to forecast consequences. This leads the AI to arrive at recommendations that deviate significantly from reality when encountering judicial scenarios outside the system. The purpose of introducing artificial intelligence systems into judicial trials is to improve efficiency and effectively assist judicial officers in handling cases, but in the current situation of poor accuracy of artificial intelligence systems, judges have to spend more time and energy to correct the errors brought about by the machines, which is more than worth the loss.

Synthetic data, based on the statistical features of real data, ensures a high level of accuracy. Unlike judicial data, which is impossible to get in real life for many complex reasons, synthetic data can be made automatically and filled in to make sure there is a variety of data, which guarantees the accuracy of the AI model[4]. The accuracy of the data is to assure the accuracy of the AI application of the basic conditions; only the accuracy of the AI can improve in order to aid the adjudicator in making accurate and fair decisions. Synthetic data "comes with a label," but also saves a lot of personnel and time costs spent on the original data labelling and processing in order to reduce the burden of the court's "many cases and few people" while also increasing the professionalism and accuracy of the data model. The enhancement of accuracy is also the improvement of efficiency, and judges do not have to dedicate too much work to non-judicial topics, which enhances the timeliness of trials and helps to ensure judicial justice in the procedure.

4.2. Mitigating the Risk of Algorithmic Discrimination

Artificial intelligence classifies, learns, and summarises enormous volumes of data by mimicking people in order to make predictions or solve problems. However, when the data and algorithms themselves lack neutrality, contain flaws, or incorporate implicit human values, AI may exhibit bias in its actual application, a phenomenon known as algorithmic discrimination[13]. And understanding the causes of prejudice and discrimination caused by AI is a vital step to improving the interpretability of judicial AI and data. Currently, judicial data reserves are incomplete and do not have a balanced distribution across many dimensions such as time, region, and case types, leaving a big data gap. The algorithm will capture the characteristics and laws of the existing court dataset as a sample, but it will discriminate against groups outside the sample, potentially leading to the absence of judicial rights protection for these groups. The quality of the training data directly influences knowledge graph formation and algorithm creation, which eventually affects the effectiveness of the AI system. Algorithms struggle to accurately identify and filter discriminatory data, and they don't pre-screen and filter this discriminatory data to facilitate their learning. The algorithms will store and deeply learn the discriminatory information, and they are very likely to output the same discriminatory results in similar scenarios, exacerbating the problem of algorithmic discrimination. There isn't enough or complete judicial data, which means that the semantic recognition function of the AI trial system and the accuracy of refining laws have a lot of room to grow. A substantial number of people, including judicial professionals, engineers, programmers, and others from commercial technology businesses, must invest in data processing to increase the efficiency of court data utilization. The involvement of many parties in the processing of real data is one of the key reasons for algorithmic bias and prejudice. The courts are inevitably influenced by subjective impressions that make it impossible to maintain neutrality, and the risk of technicians outside of the court implanting their personal beliefs and values into "labels" and algorithms is far greater. Without a legal background, it is impossible to ensure that a private company's technicians do not purposefully or inadvertently add their own value biases into the construction of knowledge graph, which in turn creates biassed algorithms that can undermine the fairness of the process and the conclusion of a trial. The algorithm functions in a de-personalised and devalued way, but the creator of the algorithm is a natural human subject with subjective value judgements, and its incorporation of value judgements in data selection and model creation is an essential reality[14]. With the constant growth of artificial intelligence technology, the increase in the complexity of the algorithm is an unavoidable tendency, which also increases the difficulty of interpreting artificial intelligence. Technology companies will not publicize the system development process due to the protection of commercial secrets, and by applying biased algorithms to the AI trial system, discrimination will continue to accumulate, and people are not controllable for the algorithm, which will form an algorithmic black box, which will ultimately lead to the public's questioning of the fairness of the judgment.

The judicial application of AI should comply with the principles of fairness and impartiality, that AI products and services need to ensure that they are free of discrimination and bias, and that they do not affect the fairness of the trial process and outcome due to technological interventions or bias in the data or models. "Only interpretable data can train AI with interpretability.[5]" Synthetic data is generated directly from real data and is completely separate from natural persons, which can make up for the structural deficiencies of real data and minimize the dependence on human labor, so the introduction of synthetic data into the AI trial system can largely avoid personal value orientation influencing the machine algorithm. Synthetic data also has the characteristics of diversity; in some areas where there are gaps in the data or the number is small, synthetic data may automatically fill in the missing data and apply labels according to the existing data and legal requirements to accomplish the compensatory effect on the data[5]. Synthetic data increases the quality of data by correcting biases in historical data, which not only improves the overall performance of AI but also better eliminates discrimination and mistakes made by developers owing to problems with the data itself.

4.3. Enhancing the Protection of Private Information

In order to ensure judicial justice, it is necessary to ensure that "the judicial system is universally recognized by the ethics of society, and the judicial process is individually recognized by the participants in judicial activities," and the disclosure of judicial data promotes the public's recognition of the judicial system and the judicial process. The public can freely access, understand, and study judicial data through the artificial intelligence system opened up by the authorities, which makes "sunshine justice" not only exist in policy documents but really put it into practice, brings the public closer to justice, and also improves the credibility of justice. But the judicial data, such as referee documents, involves personal privacy, the principle of public order and morality, the protection of minors, and many other complex issues in order to ensure that the public's right to know and maintain social stability to find a balance between the judicial organs to study the problem, to protect judicial openness and democracy, and also the requirements of judicial justice. At present, among the judgment documents published on the judgment documents network, a lot of parties' real names, home addresses, and even identity card numbers, judges, and other sensitive personal information have been leaked, and this information is only simply processed directly and is open for public inspection. At present, although the document online system uses the intelligent anonymity engine to increase the importance of sensitive information identification and shielding, the current stage of intelligent identification technology is still not mature enough. The risk of personal information being re-identified is extremely high, and if artificial intelligence fails to identify and shield sensitive information due to language recognition barriers, then the sensitive personal information will be completely exposed to the public eye. In criminal cases, many private subjects have been digging into the private information in the publicly available judgement documents to build a "criminal information database," and the public can query a person's criminal record through the "criminal information database" without authorization from the judicial authorities, which will inevitably lead to judicial disorder and undermine judicial authority. This will undoubtedly lead to disorder in the administration of justice and weaken the judiciary's authority[15]. If personal, private information is taken and utilized by persons with ulterior reasons, it is highly likely that infinite network violence and prejudice will be developed against the parties concerned, which is obviously the reverse of the ultimate goal of justice. Secondly, the development of an artificial intelligence system is an essential step in converting natural language into computer language. This work requires the conversion of both legal thinking and computer operation ability, and the judicial personnel, due to computer-level limitations, are often unable to meet the requirements of technological development, and arrangements for a large number of judicial personnel to manually process the data will exacerbate the "more cases, fewer people". Private companies have had to outsource the development of online systems for adjudication documents due to the dual problems of technology and manpower, potentially leading to the collection of large amounts of private, undisclosed information by third parties. With the emergence of generative artificial intelligence such as ChatGPT, illegal data crawling behaviour is commonplace, which poses a significant risk to personal privacy and even national data security protection. As the country's highly sensitive data, data theft, and privacy leakage have terrible consequences, the consequences are incalculable.

We can highly desensitise synthetic data to real data and process real data using data encryption, data fuzzification, and data perturbation to remove real and sensitive information. Synthetic data retains only the statistical characteristics of the original judicial data, and training AI with synthetic data can avoid the risk of privacy leakage at the initial stage. Undoubtedly, synthetic data is more effective than the prior basic anonymization method in addressing the issue of sensitive information leaking and safeguarding personal privacy. The application of synthetic data to judicial AI has made it possible for documents or evidence materials involving juvenile cases, bloody and violent cases, and other cases that are of great concern to society but are not available to the public due to privacy protection and social impact to be made available online. The scope of the court's disclosure of judicial data has been further expanded, which not only alleviates the imbalance between prosecution and defense structures and information asymmetry but also gives full play to the utility of data resources and promotes data sharing.

5. Conclusions

Artificial intelligence trial systems can help judges understand cases, analyse evidence more quickly and accurately, and provide relevant case references, effectively reducing the impact of subjective factors on judicial decisions, improving judicial impartiality, and maintaining the consistency and stability of judicial decisions. However, care should be taken in practical application to ensure that the algorithms and data of the AI system are fair, transparent, and unbiased, so as to avoid unjust judgments. The introduction of synthetic data improves the performance of AI while ensuring the data security of the AI system, and safe data ensures fair adjudication. Although AI technology can provide ancillary decision-making support, the judge's discretion remains critical. When using AI technology, judges should retain the right to make the final decision on a verdict and be able to review and evaluate the AI system's output. We must strike a dynamic balance between AI trials and judges' discretion to leverage the benefits of AI technology, enhance judicial efficiency and fairness, protect judges' professional judgement and judicial independence, and uphold the principles of judicial justice. technology in the judicial field, we must continuously explore and improve relevant laws, regulations, and ethical guidelines.

In light of the opportunities and challenges brought about by scientific and technological advancements for the judiciary, it is necessary to view artificial intelligence technology in a rational manner. Despite the risks encountered in exploring judicial intelligence, the trend of in-depth integration between artificial intelligence and judicial trials remains unstoppable. Applying emerging technologies like synthetic data in the process of building intelligent court system along with optimizing the judicial database and promoting data sharing, can help modernize the judiciary and empower it to achieve judicial justice.

References

[1] Zuo W M. Some Prospects for the Use of Legal Artificial Intelligence in China[J]. Tsinghua University Law Journal, 2018, 12(02): 108-124.

[2] Li J R. Practice and Exploration of Intelligence-controlled Trial Process Node[J]. People's Judicature, 2022, (13): 74-77+105.

[3] Cheng X Q, Chen W. Synthetic data for AI[J]. Bulletin of National Natural Science Foundation of China, 2022, 36 (03): 442-444+446.

[4] Cao J F, Chen C Y. On the AIGC Wave, Synthetic Data Is About the Future of Artificial Intelligence[J]. New Economy Leader, 2022, (04): 25-31.

[5] Du J Y, Wang H L. Synthetic Data Applications in Digital Transformation of Tax Administration[J]. Taxation Research, 2023, (07): 62-69.

[6] Pei W. Big Data with Regard to Personal Data and Criminal Due Process: Conflict and its Coordination[J]. Chinese Journal of Law, 2018, 40(02): 42-61.

[7] Bian J L, Responsibility Deconstruction and System Response of Intelligent Trial[J]. Law-Based Society, 2023, (05): 1-11.

[8] Gao X. Legal Knowledge Graph Construction for Civil Justice Applications of Artificial Intelligence[J].Law and Social Development, 2018, 24(06): 66-80.

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

[10] Gong P X, Liu M. On the Value Implications of Judicial Justice and Institutional Guarantees[J]. Studies in Law and Business, 1999, (05): 50-57.

[11] Zhang C H. A Probe of Vagueness and Its Reasons in Forensic Language[J]. Hebei Law Science, 2010, 28(09): 174-177.

[12] Wang Z. The Ways of Achieving "Quasi-syllogism" in Judicial Artificial Intelligence Ratiocination Assistance[J]. Tribune of Political Science and Law, 2022, 40(05): 28-39.

[13] Liu C. Manifestations, Causes and Governance Strategies of Algorithmic Discrimination[J]. People's Tribune, 2022, (02): 64-68.

[14] Bian J L, Qian C. The Application Limitand Procedure Regulation on the Investigations using Big Data Methods [J]. Guizhou Social Sciences, 2022, (03): 78-86.

[15] Yu Z G. Systematic Construction Regarding the Chinese Crime Record Regime: Some Thoughts about Publishing Judgment Documents Online in Current Judicial Reform[J]. Modern Law Science, 2014, 36(05): 170-184.